

Covariance regularization in inverse space

Genta Ueno^{a*} and Takashi Tsuchiya^b

^aThe Institute of Statistical Mathematics, ROIS, Tokyo, Japan / Japan Science and Technology Agency, Tokyo, Japan

^bThe Institute of Statistical Mathematics, ROIS, Tokyo, Japan

Abstract: In data assimilation, covariance matrices are introduced in order to prescribe the weights of the initial state, model dynamics, and observation, and suitable specification of the covariances is known to be essential for obtaining sensible state estimates. The covariance matrices are specified by sample covariances and are converted according to an assumed covariance structure. Modeling of the covariance structure consists of the regularization of a sample covariance and the constraint of a dynamic relationship. Regularization is required for converting the singular sample covariance into a non-singular sample covariance, removing spurious correlation between variables at distant points, and reducing the required number of parameters that specify the covariances. In previous studies, regularization of sample covariances has been carried out in physical (grid) space, spectral space, and wavelet space. We herein propose a method for covariance regularization in inverse space, in which we use the covariance selection model (the Gaussian graphical model). For each variable, we assume neighboring variables, i.e., a targeted variable is directly related to its neighbors and is conditionally independent of the non-neighboring variables. Conditional independence is expressed by specifying zero elements in the inverse covariance matrix. The non-zero elements are estimated numerically by the maximum likelihood using Newton's method. Appropriate neighbors can be selected with the AIC or BIC information criteria. We address some techniques for implementation when the covariance matrix has a large dimension. We present an illustrative example using a simple 3×3 matrix and an application to a sample covariance obtained from sea surface height (SSH) observations.

KEY WORDS data assimilation; 4D-Var; ensemble Kalman filter; covariance selection; Gaussian graphical model

1 Introduction

In data assimilation, covariance matrices are introduced in order to prescribe the properties of the initial state, the system noise, and the observation noise. The inverse matrices of the respective covariances work as weights of initial estimates, model dynamics, and observations. Suitable specification of the covariance matrices is essential for obtaining sensible estimates, and misspecification of the matrices may lead to overfitting or underfitting of data and/or failure of the assimilation altogether (e.g., Fukumori, 2001). The present paper deals with a technique for building covariance matrices in data assimilation.

Figure 1 shows a procedure for specifying a covariance matrix in data assimilation. The specification process of covariance matrices consists of two tasks: (1) calculation of statistics (sample covariance) and (2) modeling of the covariance structure. From this perspective, we would like to review techniques for covariance specification that have been reported previously. In the present paper, we propose a technique that contributes to the modeling of the covariance structure.

Task 1: Calculation of statistics (sample covariance)

- Model output/ensemble
- Observation (raw, scaled, detrended)
- Difference between model and observation

Task 2: Modeling of the covariance structure

- Regularization
 1. Physical space
 - Diagonal (independence)/Block diagonal
 - Correlation models
 - Digital filters
 - Compact support/moderation function
 2. Spectral space
 - Harmonics expansion (isotropy)
 - Diagonal/Block diagonal
 - Compact support
 3. Wavelet space
 - Diagonal (local average)
 - Diagonal + some off-diagonal
 4. Inverse space (**Present study**)
 - Diagonal + some off-diagonal (**conditional independence**)
- Dynamic constraints

Figure 1. Specification of a covariance matrix in data assimilation.

*Correspondence to: The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan. E-mail: gen@ism.ac.jp

1.1 Calculation of statistics (sample covariance)

The first task in covariance specification is to calculate a statistic, that is, to calculate the sample covariance matrix of a model output, observations, or some quantity computed from these values. Model output is often used to specify a covariance matrix that represents the uncertainty of the initial state, which is known as a background error covariance matrix in the literature on variational assimilation. The NMC method (Parrish and Derber, 1992) has used a pair of model estimates based on different lead times to model the background error covariance. With a single model run, the CQ method (Polavarapu *et al.*, 2005; Jackson *et al.*, 2008) calculates a sample covariance of the difference between the model outputs separated by a fixed time. Model ensemble is also used to compute a sample covariance matrix, in which the ensemble members are generated by assimilating perturbed observation or imposing perturbed model parameters or initial/boundary conditions. (See Houtekamer *et al.* (1996); Fisher (2003); Buehner (2005); Borovikov *et al.* (2005); Alves and Robert (2005) for background error covariances and Trémolet (2007) for system noise (model error) covariances.)

Regarding observation noise covariances, sample covariance matrices are computed from the data, from the data multiplied by a scale factor (Oke *et al.*, 2002), or from the detrend data (Ueno *et al.*, 2007). The difference between model outputs and observations is used in the covariance matching method (Fu *et al.*, 1993) and its extended version (Menemenlis and Chechelnitsky, 2000), which specifies the covariance matrices for system noise and observation noise simultaneously for a linear system and observation equations.

The sample covariance matrices, however, cannot be used as they are in many applications, mainly due to the following three reasons. The first reason is mathematical. That is, the sample covariance matrices are usually singular because the number of ensemble members or the number of time steps is small compared to the number of state variables or the number of observation sites. Due to the imbalance between the dimension and the sample size, calculated sample covariance matrices become rank-deficient, and therefore singular. In practice, the singular covariance matrices results in some inconvenience. For example, since the inverse matrices do not exist, neither the background term nor the observation term can be evaluated in cost functions.

The second reason is that the sample covariances contain spurious correlations between distant grid points. The correlations are considered to be physically inconsistent, again due to the limited number of samples. The third reason is practical and is related to memory capacity. As a matter of course, a covariance matrix has elements of the square of the number of variables, which may amount to 10^{14} in a background error covariance, for example. The size of the required memories makes the data assimilation difficult or impossible due to the huge computational cost.

1.2 Modeling of the covariance structure

These three difficulties in sample covariance matrices motivate the second task in covariance specification, namely, modeling of the covariance structure. In this task, we impose a regularization condition and/or a dynamic constraint to the covariance.

1.2.1 Regularization

Regularization is an attempt to convert a sample covariance matrix into another matrix that is non-singular, does not suffer from spurious correlations, and can be specified with a smaller number of parameters than the squared number of variables. At the same time, the matrix is expected to maintain to some degree the properties of the original sample covariance. We briefly review the regularization procedures, focusing on the space in which the regularization procedures are carried out: (1) physical (grid) space, (2) spectral (Fourier) space, and (3) wavelet space.

Physical space The simplest way to regularize a covariance matrix is to assume the matrix to be diagonal in physical space. That is, off-diagonal elements of the sample covariance are forced to be zero. The diagonalized matrix is apparently non-singular and is represented by parameters of dimension equal to the number of variables. This procedure is widely adopted in many data assimilation experiments. The diagonal assumption, however, not only discards the spurious correlations for distant points, but also neglects the realistic correlation structure for close points of model grids and observation sites.

The second way to regularize a covariance matrix is to assume the matrix to have elements represented by a parameterized analytic function that is known to give positive definite covariance. Such a function can be selected from covariance function families that are known in geostatistics, such as the Matérn family (e.g., Diggle and Ribeiro Jr., 2007). In data assimilation, a Gaussian function and autoregressive functions are well-known correlation functions (e.g., Daley, 1991, p.117). Frehlich (2006) used structure functions to construct a spatial-varying observation noise matrix.

Digital filters can be applied to build a covariance that approximates a Gaussian. Iterative Laplacian filters (Derber and Rosati, 1989; Egbert *et al.*, 1994; Weaver and Courtier, 2001) and Gaussian recursive filters (Lorenc, 1992, 1997; Purser *et al.*, 2003a,b) have been developed. Digital-filter-based correlations are well suited for ocean models that have complex lateral boundary conditions, while atmospheric models that have no lateral boundaries often adopt the spherical harmonic expansion, as described later herein. In recent studies, digital filters have been used to dampen high-frequency short-scale oscillations that are present in sample covariances of very small ensembles (Keppenne *et al.*, 2008; Raynaud *et al.*, 2008).

Compactly supported correlation functions, which represents correlations essentially vanishing beyond

a certain cut-off distance, are used to remove spurious long-range correlations (Gaspari and Cohn, 1999; Gaspari *et al.*, 2006; Gneiting, 1999, 2002). Compactly supported correlation functions are often used in an ensemble Kalman filter (EnKF, Evensen, 2003) implementation, by taking the Schur product (the Hadamard product) with ensemble covariances (Houtekamer and Mitchell, 2001; Hamill *et al.*, 2001). Flow-dependent moderation functions (Bishop and Hodyss, 2007) have been proposed.

Spectral space The second space for the regularization is the spectral space. Hollingsworth and Lönnerberg (1986) and Lönnerberg and Hollingsworth (1986) have modeled correlation functions in terms of cylindrical harmonics (Fourier-Bessel expansion, (e.g., Daley, 1991, Appendix G)), which was adopted for the local analysis of correlation functions based on the tangent plane to the sphere. On the sphere itself, spherical harmonics (Legendre expansion) was used to model isotropic correlation functions. With a diagonal matrix in spectral space, Parrish and Derber (1992) represented isotropic correlations for observation noise. Courtier *et al.* (1998) used a block-diagonal matrix to model background error covariances, where the vertical correlation scale depends on the horizontal scale. The compactly supported correlation functions were also used in spectral space. Buehner and Charron (2007) pointed out the importance of spectral localization in the reduction of sampling error.

Wavelet space In the third regularization space, wavelet space, regularization in physical space and spectral space can be carried out simultaneously with wavelets, which are localized in both physical space and spectral space. In wavelet space, sample covariances were regularized with diagonal covariance matrices (Fisher and Andersson, 2001; Deckmyn and Berre, 2005) and matrices having diagonal elements and some off-diagonal elements (Nychka *et al.*, 2002; Rhodin and Anlauf, 2007). Similar to compactly supported correlation functions, wavelets can filter out spurious noise on sample covariances (Pannekoucke *et al.*, 2007).

1.2.2 Dynamic constraints

In addition to the covariance regularization, imposing dynamic constraints is another procedure of the covariance structure modeling. Based on the semigeostrophic theory, Desroziers (1997) has assumed isotropic correlations in geostrophic space and has transformed the coordinates into real space to obtain anisotropic and flow-dependent correlations. The balance operator has been employed to transform a horizontally homogeneous and isotropic covariance for the unbalanced part of the variables into a covariance of the variables, including the effects of the balanced part of the variables in accordance with the dynamic relationship (Gauthier *et al.*, 1998; Derber and Bouttier, 1999; Cullen, 2003). Riishøjgaard (1998) has introduced a field term, which is a matrix in

which the elements depend on the difference between the background state estimates, so that flow-dependency is taken into account.

1.3 Present study: Regularization in inverse space

In the present paper, we propose an alternative method for the regularization of covariance matrices in the fourth space: inverse space. Based on the statistics, we model an inverse matrix to obtain a valid covariance matrix. As described below, the inverse matrices are related to the concept of the conditional independence of variables. This concept contrasts modeling in physical, spectral, and wavelet space, where no concept of probability was included. Assumptions on an inverse matrix can be found in a study by Chin *et al.* (1999), who used a Markov random field to truncate the inverse covariance matrix of the Kalman filter.

In Section 2, we review a sample covariance matrix, which gives a basic statistic of the covariance matrix, and its relation to the maximum likelihood estimator (MLE) of the covariance matrix of the assumed Gaussian distribution. To overcome the rank deficiency of the sample covariance, we parametrize the elements of an inverse matrix using the concept of a covariance selection model (Dempster, 1972), which has recently come to be known as a Gaussian graphical model (Lauritzen, 1996), as described in Section 3. In Section 4, we present a simple example using a 3×3 covariance matrix. We address some implementation techniques that are useful especially for large problems in Section 5. Section 6 shows an application to a sample covariance matrix obtained from sea surface height (SSH) observation, which may be used as an observation noise matrix. In Section 7, we discuss the properties of the regularized covariances in inverse space, and conclusions are presented in Section 8.

2 Sample covariance

Let us review the basic concept of sample covariance matrices. When n -dimensional vector data $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{N_e}$ are available, the sample mean and the sample covariance are defined, respectively, as

$$\bar{\mathbf{x}} = \frac{1}{N_e} \sum_{i=1}^{N_e} \mathbf{x}_i, \quad (1)$$

$$\mathbf{S} = \frac{1}{N_e} \sum_{i=1}^{N_e} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})', \quad (2)$$

where the prime (') denotes the transpose.

The sample mean and the sample covariance are known as the MLEs of the mean and covariance of a Gaussian distribution. Consider a Gaussian distribution of mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$:

$$p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}} \exp \left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right], \quad (3)$$

where $|\Sigma|$ denotes the determinant of Σ . When data $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{N_e}$ are given, the log-likelihood becomes as follows (see Appendix A):

$$\begin{aligned} \ell(\boldsymbol{\mu}, \Sigma) &= \sum_{i=1}^{N_e} \log p(\mathbf{x}_i; \boldsymbol{\mu}, \Sigma) \\ &= \sum_{i=1}^{N_e} \left[-\frac{n}{2} \log 2\pi - \frac{1}{2} \log |\Sigma| \right. \\ &\quad \left. - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right] \\ &= -\frac{N_e}{2} [n \log 2\pi - \log |\Sigma^{-1}| + \text{tr} \mathbf{S} \Sigma^{-1} \\ &\quad + (\bar{\mathbf{x}} - \boldsymbol{\mu})' \Sigma^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu})], \end{aligned} \quad (4) \quad (5) \quad (6)$$

where tr denotes the trace operator. The MLEs of $\boldsymbol{\mu}$ and Σ are the sample mean and sample covariance, respectively:

$$\bar{\boldsymbol{\mu}} = \bar{\mathbf{x}}, \quad (7)$$

$$\bar{\Sigma} = \mathbf{S}, \quad (8)$$

if $N_e \geq n + 1$ (e.g., Magnus and Neudecker, 1999). The inequality is a necessary condition for the Hessian matrix of the log-likelihood given by Eq. (6) to be negative definite, which confirms that $(\bar{\boldsymbol{\mu}}, \bar{\Sigma})$ is a strict local maximum. In the case of $N_e \leq n$, however, although $\bar{\mathbf{x}}$ and \mathbf{S} satisfy the stationary point condition (the gradient vector becomes zero), the Hessian matrix becomes singular, and consequently we cannot tell whether $(\bar{\mathbf{x}}, \mathbf{S})$ is a maximum point, a minimum point, or a saddle point.

3 Regularization of sample covariance in inverse space

3.1 Covariance selection models

As described in the previous section, the sample covariance \mathbf{S} (Eq. 2) is not necessary an MLE of the covariance Σ when $N_e \leq n$. In addition, \mathbf{S} becomes singular because \mathbf{S} is an $n \times n$ matrix but has a rank of at most $N_e - 1$. These undesirable properties appear because the number of free parameters of the covariance model is too large compared with the number of samples.

To circumvent these difficulties, we assume a structure on Σ to reduce its degree of freedom, so that we can estimate Σ from small samples. Specifically, we assume a sparse structure in the inverse space of Σ , not in the usual physical (grid) space, by assigning zeros for most elements of Σ^{-1} . In statistics, models with sparse inverse covariances are called covariance selection models or Gaussian graphical models (Dempster, 1972; Lauritzen, 1996). A simple example for three variables is shown in Figure 2a. In this example, σ^{13} and σ^{31} are set to zero, where σ^{ij} denotes the (i, j) -elements of Σ^{-1} .

Zero elements in Σ^{-1} are equivalent to the conditional independence of the corresponding pairs of variables given the remaining variables (e.g., Lauritzen, 1996, Proposition 5.2). Conditional independence is naturally

(a) Inverse covariance matrix (b) Undirected graph

$$\Sigma^{-1} = \begin{pmatrix} \sigma^{11} & \sigma^{12} & 0 \\ \sigma^{12} & \sigma^{22} & \sigma^{23} \\ 0 & \sigma^{23} & \sigma^{33} \end{pmatrix} \quad \begin{array}{ccc} \circ & \circ & \circ \\ 1 & 2 & 3 \end{array}$$

Figure 2. Representation of conditional independence between variables 1 and 3 given variable 2 as (a) an inverse covariance matrix and as (b) an undirected graph.

associated with an undirected graph in which the nodes correspond to the variables. An edge in the graph represents the interaction between the two nodes that the edge connects, and no edge represents the conditional independence between the two variables given the other variables. Graphs constructed in this manner are referred to as *independence graphs* or *interaction graphs*. As an example, the inverse covariance shown in Figure 2a is equivalent to the conditional independence between variables 1 and 3 given variable 2. The conditional independence is displayed by the independence graph shown in Figure 2b, which has two edges between node 1 and node 2 and between node 2 and node 3 but has no edge between node 1 and node 3.

The independence graph captures causal relations among variables. We explain this characteristic using an example to clarify the meaning of the model. Suppose that variable 1 represents the sales of fans, variable 2 represents temperature, and variable 3 represents air-conditioner sales. Correlation appears to exist between variables 1 and 3. However, a real causal relation exists between variables 2 and 1 and between variables 2 and 3. The correlation between variable 1 and variable 3 is implied by the two causal relations. The aforementioned example of a covariance selection model (Figure 2b) is suitable for explaining this situation. In the independence graph, while we have two edges between 1 and 2 and between 3 and 2, there is no edge connecting 1 and 3. This structure indicates exactly what we explained above.

The remainder of the non-zero elements in Σ^{-1} are estimated using the maximum likelihood method to obtain $\hat{\Sigma}^{-1}$ and consequently $\hat{\Sigma}$. It is known that the maximum likelihood estimate $\hat{\Sigma}^{-1}$ has the following remarkable characteristics. That is, the covariance in the physical space $\hat{\Sigma} \equiv (\hat{\Sigma}^{-1})^{-1}$ is positive definite (and, therefore, non-singular) and has elements that are identical to those of the original sample covariance \mathbf{S} in positions where the non-zero elements were assumed in Σ^{-1} (e.g., Lauritzen, 1996, Theorem 5.3). In the three-dimensional example with $\sigma^{13} = \sigma^{31} = 0$ (Figure 2a), $\hat{\Sigma}$ has elements identical to \mathbf{S} in the positions $(1, 1)$, $(1, 2)$, $(2, 1)$, $(2, 2)$, $(2, 3)$, $(3, 2)$, and $(3, 3)$, which are consistent with positions at which non-zero elements were assumed in Σ^{-1} .

3.2 Specification of non-zero elements in the inverse

Now, we propose to apply a covariance selection model for the regularization of sample covariances. This is justified as follows. Since variables are defined at grid points

in physical space, it is natural to assume local interactions between neighboring variables that are located at close points. If the nonzero pattern of Σ^{-1} reflects the neighbor structure of the grid points, the covariance selection models exactly meets such a requirement.

We construct a covariance selection model Σ^{-1} by specifying non-zero elements. Before examining local interaction of variables, we can require that all of the diagonal elements of Σ^{-1} be positive, which is a result of the positive definiteness of Σ . If we stop here, the diagonal inverse Σ^{-1} leads to a diagonal covariance Σ , which means that each pair of variables is independent and equivalent to a diagonal assumption in physical space. This condition can be displayed with a graph consisting of nodes alone with no edges. Examination of one such node reveals that it has no edges, as shown in Figure 3a.

Off-diagonal elements of Σ^{-1} can be introduced based on the spatial arrangement of variables. That is, conditional independence can be modeled by an independence graph having nodes that are located at grid points at which variables are defined. Assuming a two-dimensional grid space, for example, we may naturally assume local interaction with four variables located at the nearest grid points in two directions (the zonal and the meridional directions). This relationship is represented by a graph having four edges from the centered node to four adjacent nodes (Figure 3b). The configuration also indicates the conditional independence of the variables outside the four adjacent variables. The inverse covariance Σ^{-1} has non-zero elements with regard to the four neighboring variables and zero elements with more distance variables.

When we include interaction with variables located at oblique grid points, the oblique nodes also become neighbors (that is, connected by edges to the central node) as in Figure 3c. As before, a variable is permitted to interact with the surrounding eight variables and is conditionally independent of the variables outside of the eight variables. More structured models may be assumed by activating edges between nodes, as shown in Figures 3d–3h, in a similar manner.

In any graph configuration, once we assume non-zero elements in Σ^{-1} , we can write the inverse Σ^{-1} as a linear combination of fixed matrices as a function of their coefficients:

$$\Sigma^{-1}(\beta) = \sum_{k=1}^m \mathbf{A}_k \beta_k, \quad (9)$$

where \mathbf{A}_k ($k = 1, \dots, m$) is a fixed matrix that indicates non-zero elements for interacting pairs of variables, β_k ($k = 1, \dots, m$) is a scalar coefficient to be estimated, and β denotes a vector composed by a set of coefficients, $(\beta_1 \dots \beta_m)'$. Specifically, \mathbf{A}_k ($k = 1, \dots, m$) can be taken as a set of matrices, each of which has one diagonal element with a value of unity, or has two off-diagonal elements with value of unity for an interacting pair. Instead of unity, we may assume, without loss of generality, that the values of the diagonal elements are two. The latter specification is convenient for constructing an efficient algorithm (see Section 5).

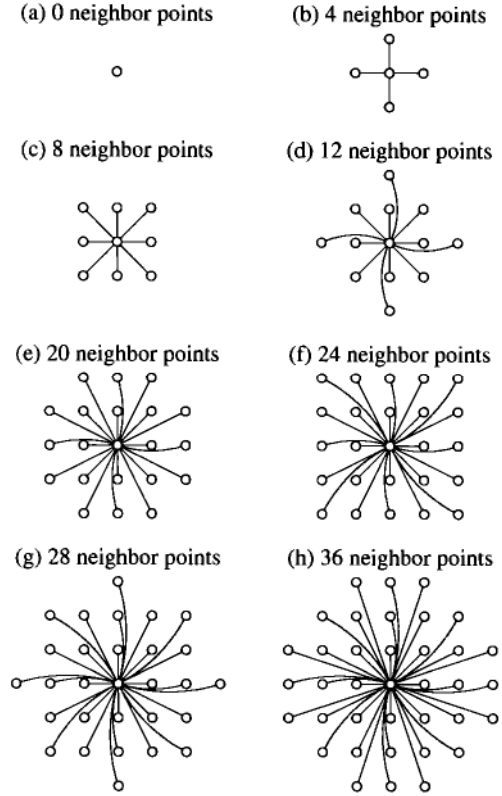


Figure 3. Graphical representation of dependence between variables. For each panel, the central node corresponds to a variable for which the direct relationship (denoted by edges) with the other variables is considered. Figure (a) shows a variable that is independent of the other variables. The other figures show variables that interact with their adjacent variables located at (b) four, (c) eight, (d) 12, (e) 20, (f) 24, (g) 28, and (h) 36 neighbor points, respectively.

3.3 Maximum likelihood estimation of the parameters

Under the condition that $\Sigma^{-1} = \Sigma^{-1}(\beta)$ as in Eq. (9), we estimate μ and β by maximizing the log-likelihood

$$\begin{aligned} \ell(\mu, \beta) = & -\frac{N_e}{2} [n \log 2\pi - \log |\Sigma^{-1}(\beta)| \\ & + \text{tr } S \Sigma^{-1}(\beta) \\ & + (\bar{\mathbf{x}} - \mu)' \Sigma^{-1}(\beta) (\bar{\mathbf{x}} - \mu)] \end{aligned} \quad (10)$$

The stationary point of μ is again

$$\hat{\mu} = \bar{\mathbf{x}}, \quad (11)$$

and also becomes the MLE if Σ is non-singular. Regarding β , we cannot obtain the stationary points analytically. We then numerically search optimal β that maximizes

$$\begin{aligned} \ell(\bar{\mathbf{x}}, \beta) = & -\frac{N_e}{2} [n \log 2\pi - \log |\Sigma^{-1}(\beta)| \\ & + \text{tr } S \Sigma^{-1}(\beta)]. \end{aligned} \quad (12)$$

Maximizing the log-likelihood given by Eq. (12) is equivalent to minimizing a cost function, as shown below:

$$f(\beta) = \text{tr } S \Sigma^{-1}(\beta) - \log |\Sigma^{-1}(\beta)|. \quad (13)$$

We find a solution $\hat{\beta}$ that minimizes $f(\beta)$ with Newton's method, because the gradient vector and the Hessian matrix of the cost function given by Eq. (13) are obtained analytically when $\Sigma^{-1}(\beta)$ is parameterized as Eq. (9). The Newton direction (or the Newton step) $\Delta\beta$ is given by

$$\Delta\beta = -[\nabla^2 f(\beta)]^{-1} \nabla f(\beta). \quad (14)$$

The gradient vector $\nabla f(\beta) = (\partial f / \partial \beta_i)$ and the Hessian matrix $\nabla^2 f(\beta) = (\partial^2 f / \partial \beta_i \partial \beta_j)$ are obtained as

$$\frac{\partial f}{\partial \beta_i} = \text{tr} \mathbf{S} \mathbf{A}_i - \text{tr} \Sigma(\beta) \mathbf{A}_i, \quad (15)$$

$$\frac{\partial^2 f}{\partial \beta_i \partial \beta_j} = \text{tr} \Sigma(\beta) \mathbf{A}_i \Sigma(\beta) \mathbf{A}_j. \quad (16)$$

Their derivations are given in Appendices B.1 and B.2 [Eqs. (104) and (119)].

Note that the maximum likelihood estimation is formulated as a convex optimization problem (e.g., Boyd and Vandenberghe, 2004, Chapter 4), and hence can be computed efficiently and rigorously using Newton's method by maximizing (globally) the likelihood function. Recent convex optimization theory shows that this optimization problem can be solved in polynomial-time (Vandenberghe *et al.*, 1998; Boyd and Vandenberghe, 2004; Tsuchiya and Xia, 2007).

3.4 Selection of the number of neighbors

As shown in Figure 3, several covariance selection models may be applied to obtain a regularized covariance. Since one regularized covariance is sufficient to conduct a data assimilation experiment, we select a preferable covariance among those estimated by different models. One possible method by which to objectively select a preferable number of neighbors is to use information criteria (e.g., Konishi and Kitagawa, 2007). We select the number of coefficients $m = \dim \beta$ using the Akaike information criterion (AIC, Akaike, 1974)

$$\text{AIC} = -2\ell(\bar{\mathbf{x}}, \hat{\beta}) + 2m, \quad (17)$$

or the Bayesian information criterion (BIC, Akaike, 1977; Schwarz, 1978) defined as

$$\text{BIC} = -2\ell(\bar{\mathbf{x}}, \hat{\beta}) + m \log N_e, \quad (18)$$

where

$$\hat{\beta} = \underset{\beta}{\text{argmax}} \ell(\bar{\mathbf{x}}, \beta). \quad (19)$$

Among the maximum likelihood models of different numbers of neighbors, we can select the best model by finding the model with the smallest AIC or that with the smallest BIC.

4 Illustrative example

4.1 Analytical consideration

Let us consider the simplest example, a 3×3 sample covariance matrix:

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & s_{13} \\ s_{12} & s_{22} & s_{23} \\ s_{13} & s_{23} & s_{33} \end{pmatrix}. \quad (20)$$

The sample covariance \mathbf{S} is assumed to be computed from $N_e (\leq 3)$ samples and is therefore singular:

$$|\mathbf{S}| = s_{11}s_{22}s_{33} + 2s_{12}s_{13}s_{23} - s_{11}s_{23}^2 - s_{22}s_{13}^2 - s_{33}s_{12}^2 \quad (21)$$

$$= 0. \quad (22)$$

It would be desirable to convert \mathbf{S} into a regularized matrix Σ .

We consider a covariance selection model shown in Figure 2, in which variables 1 and 3 are conditionally independent given variable 2, and a regularized inverse Σ^{-1} has zeros at (1, 3)- and (3, 1)-elements:

$$\Sigma^{-1}(\beta) = \begin{pmatrix} \beta_1 & \beta_4 & \\ \beta_4 & \beta_2 & \beta_5 \\ & \beta_5 & \beta_3 \end{pmatrix}, \quad (23)$$

where $\beta = (\beta_1 \ \beta_2 \ \beta_3 \ \beta_4 \ \beta_5)'$ and blanks denote zero elements. As noted in Eq. (9), the regularized inverse $\Sigma^{-1}(\beta)$ is represented as a linear combination of five matrices $\mathbf{A}_1, \dots, \mathbf{A}_5$:

$$\Sigma^{-1}(\beta) = \sum_{k=1}^5 \mathbf{A}_k \beta_k, \quad (24)$$

where

$$\begin{aligned} \mathbf{A}_1 &= \begin{pmatrix} 1 & & \\ & & \\ & & \end{pmatrix}, & \mathbf{A}_2 &= \begin{pmatrix} & 1 & \\ & & \\ & & \end{pmatrix}, \\ \mathbf{A}_3 &= \begin{pmatrix} & & 1 \\ & & \\ 1 & & \end{pmatrix}, & \mathbf{A}_4 &= \begin{pmatrix} & 1 & \\ 1 & & \\ & & \end{pmatrix}, \\ \mathbf{A}_5 &= \begin{pmatrix} & & \\ & 1 & \\ 1 & & \end{pmatrix}. \end{aligned} \quad (25)$$

The log-likelihood given by Eq. (12) becomes

$$\ell(\bar{\mathbf{x}}, \beta) = -\frac{N_e}{2} [3 \log 2\pi - \log |\Sigma^{-1}(\beta)| + \text{tr} \mathbf{S} \Sigma^{-1}(\beta)] \quad (26)$$

The optimality condition $\partial \ell / \partial \beta_i = 0$ ($i = 1, \dots, 5$) is equivalent to (see Eq. (15))

$$\text{tr} \mathbf{S} \mathbf{A}_i - \text{tr} \Sigma(\beta) \mathbf{A}_i = 0, \quad (27)$$

for $i = 1, \dots, 5$. Substituting Eq. (25), we obtain

$$s_{11} - \sigma_{11} = 0 \quad (28)$$

$$s_{22} - \sigma_{22} = 0 \quad (29)$$

$$s_{33} - \sigma_{33} = 0 \quad (30)$$

$$2s_{12} - 2\sigma_{12} = 0 \quad (31)$$

$$2s_{23} - 2\sigma_{23} = 0 \quad (32)$$

where σ_{ij} denotes the (i, j) -element of $\Sigma(\beta)$. That is, if there exist coefficients β that satisfy the optimality condition, then the constructed covariance matrix $\Sigma(\beta)$ has values that are identical to the sample covariance S in the $(1, 1)$ -, $(2, 2)$ -, $(3, 3)$ -, $(1, 2)$ -, and $(2, 3)$ - elements: $\sigma_{11} = s_{11}$, $\sigma_{22} = s_{22}$, $\sigma_{33} = s_{33}$, $\sigma_{12} = s_{12}$, and $\sigma_{23} = s_{23}$. Note that the locations of these elements are identical to those of the elements of $\Sigma^{-1}(\beta)$, where non-zeros were imposed.

Since $\Sigma(\beta)$ is the inverse of $\Sigma^{-1}(\beta)$, the σ_{ij} elements are required to be expressed in terms of β as

$$\sigma_{11} = (\beta_2\beta_3 - \beta_5^2) / |\Sigma^{-1}(\beta)|, \quad (33)$$

$$\sigma_{22} = \beta_1\beta_3 / |\Sigma^{-1}(\beta)|, \quad (34)$$

$$\sigma_{33} = (\beta_1\beta_2 - \beta_4^2) / |\Sigma^{-1}(\beta)|, \quad (35)$$

$$\sigma_{12} = -\beta_3\beta_4 / |\Sigma^{-1}(\beta)|, \quad (36)$$

$$\sigma_{23} = -\beta_1\beta_5 / |\Sigma^{-1}(\beta)|, \quad (37)$$

$$\sigma_{13} = \beta_4\beta_5 / |\Sigma^{-1}(\beta)|, \quad (38)$$

where

$$|\Sigma^{-1}(\beta)| = \beta_1\beta_2\beta_3 - \beta_1\beta_5^2 - \beta_3\beta_4^2. \quad (39)$$

Equations (33) through (38) are not independent, because the equations include only five independent variables, β_1, \dots, β_5 . In fact, σ_{13} is expressed in terms of σ_{12} , σ_{23} , and σ_{22} :

$$\sigma_{13} = \beta_4\beta_5 / |\Sigma^{-1}(\beta)| \quad (40)$$

$$= \frac{\beta_4}{|\Sigma^{-1}(\beta)|} \frac{\beta_5}{|\Sigma^{-1}(\beta)|} |\Sigma^{-1}(\beta)| \quad (41)$$

$$= \left(-\frac{\sigma_{12}}{\beta_3} \right) \left(-\frac{\sigma_{23}}{\beta_1} \right) |\Sigma^{-1}(\beta)| \quad (42)$$

$$= \sigma_{12}\sigma_{23} \frac{|\Sigma^{-1}(\beta)|}{\beta_1\beta_3} \quad (43)$$

$$= \frac{\sigma_{12}\sigma_{23}}{\sigma_{22}} \quad (44)$$

This means that the assumed inverse structure given by Eq. (23) expresses a covariance of the form

$$\Sigma(\beta) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{12}\sigma_{23}/\sigma_{22} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{12}\sigma_{23}/\sigma_{22} & \sigma_{23} & \sigma_{33} \end{pmatrix}. \quad (45)$$

If the optimality condition is satisfied, combining Eqs. (28) through (32) and Eq. (45) yields the MLE of the covariance

$$\hat{\Sigma} = \begin{pmatrix} s_{11} & s_{12} & s_{12}s_{23}/s_{22} \\ s_{12} & s_{22} & s_{23} \\ s_{12}s_{23}/s_{22} & s_{23} & s_{33} \end{pmatrix}, \quad (46)$$

under the condition that coefficients β exist such that

$$\begin{pmatrix} \beta_1 & \beta_4 & \\ \beta_4 & \beta_2 & \beta_5 \\ & \beta_5 & \beta_3 \end{pmatrix} = \hat{\Sigma}^{-1}. \quad (47)$$

To satisfy the condition given by Eq. (47), the determinant of $\hat{\Sigma}$ should not vanish:

$$|\hat{\Sigma}| = s_{11}s_{22}s_{33} + \frac{(s_{12}s_{23})^2}{s_{22}} - s_{11}s_{23}^2 - s_{12}^2s_{33} \quad (48)$$

$$\neq 0. \quad (49)$$

This condition may hold even when the determinant of the sample covariance vanishes, as shown by Eq. (22).

Let us consider the case in which we further assume a matrix

$$\mathbf{A}_6 = \begin{pmatrix} & & 1 \\ & & \\ 1 & & \end{pmatrix}, \quad (50)$$

with which we parameterize Σ^{-1} as a full matrix:

$$\Sigma^{-1}(\beta) = \begin{pmatrix} \beta_1 & \beta_4 & \beta_6 \\ \beta_4 & \beta_2 & \beta_5 \\ \beta_6 & \beta_5 & \beta_3 \end{pmatrix} \quad (51)$$

using six parameters

$$\beta = (\beta_1 \ \beta_2 \ \beta_3 \ \beta_4 \ \beta_5 \ \beta_6)'. \quad (52)$$

The optimality condition yields an additional equation for the $(1, 3)$ -element

$$2s_{13} - 2\sigma_{13} = 0 \quad (53)$$

as well as Eqs. (28)–(32), which indicate that $\hat{\Sigma} = S$. The six parameters in β yield six independent equations between the elements of Σ and β :

$$\sigma_{11} = (\beta_2\beta_3 - \beta_5^2) / |\Sigma^{-1}(\beta)|, \quad (54)$$

$$\sigma_{22} = (\beta_1\beta_3 - \beta_6^2) / |\Sigma^{-1}(\beta)|, \quad (55)$$

$$\sigma_{33} = (\beta_1\beta_2 - \beta_4^2) / |\Sigma^{-1}(\beta)|, \quad (56)$$

$$\sigma_{12} = (-\beta_3\beta_4 + \beta_5\beta_6) / |\Sigma^{-1}(\beta)|, \quad (57)$$

$$\sigma_{23} = (-\beta_1\beta_5 + \beta_4\beta_6) / |\Sigma^{-1}(\beta)|, \quad (58)$$

$$\sigma_{13} = (\beta_4\beta_5 - \beta_2\beta_6) / |\Sigma^{-1}(\beta)|, \quad (59)$$

where

$$|\Sigma^{-1}(\beta)| = \beta_1\beta_2\beta_3 + 2\beta_4\beta_5\beta_6 - \beta_1\beta_5^2 - \beta_2\beta_6^2 - \beta_3\beta_4^2. \quad (60)$$

When we combine the two conditions, the existence of β requires that $|\hat{\Sigma}| = |S| \neq 0$. The requirement, however, cannot be satisfied due to the original assumption given in Eq. (22), which originates from the rank deficiency of S . Therefore, it is essential to impose zero elements in Σ^{-1} , which is equivalent to assuming conditional independence, in order to surmount the problem of rank deficiency and construct a regularized matrix.

4.2 Numerical example

We use a numerical example of a sample covariance that is singular, and demonstrate the entire process of regularization. We consider a 3×3 sample covariance

$$S = \begin{pmatrix} 0.884 & 0.780 & 0.028 \\ 0.780 & 0.876 & 0.458 \\ 0.028 & 0.458 & 1.005 \end{pmatrix}, \quad (61)$$

which is computed from three samples:

$$\mathbf{x}_1 = (0.573 \quad 0.223 \quad -1.366)', \quad (62)$$

$$\mathbf{x}_2 = (0.190 \quad 0.930 \quad 1.042)', \quad (63)$$

$$\mathbf{x}_3 = (-1.585 \quad -1.312 \quad -0.578)', \quad (64)$$

and the sample mean:

$$\bar{\mathbf{x}} = (-0.274 \quad -0.053 \quad -0.301)' \quad (65)$$

Since the rank of S is two (which is equal to the number of samples minus one), S is rank deficient and

$$|S| = 0.000. \quad (66)$$

If we assume the inverse of the form of Eq. (23) and estimate $\beta_{1:5}$ by maximizing the log-likelihood given by Eq. (26), we obtain the inverse

$$\hat{\Sigma}^{-1} = \begin{pmatrix} 5.293 & -4.715 & 0.000 \\ -4.715 & 5.700 & -0.684 \\ 0.000 & -0.684 & 1.307 \end{pmatrix}. \quad (67)$$

Since $\hat{\Sigma}^{-1}$ is found to be regular ($|\hat{\Sigma}^{-1}| = 7.900$), we can invert Eq. (67) to obtain a regularized covariance

$$\hat{\Sigma} = \begin{pmatrix} 0.884 & 0.780 & 0.408 \\ 0.780 & 0.876 & 0.458 \\ 0.408 & 0.458 & 1.005 \end{pmatrix}, \quad (68)$$

which has a finite determinant: $|\hat{\Sigma}| = 0.127$.

Table 1 summarizes regularized matrices using different models. Model (a) assumes that the three variables are independent, which leads to a diagonal inverse matrix. Three parameters, $\beta_{1:3}$, are assigned to express the matrix. Models (b) through (d) contain one edge between the nodes. Variables 1 and 2 are independent of variable 3 in model (b), and similar assumptions are made for models (c) and (d). Corresponding inverse matrices are block-diagonal, having four parameters, $\beta_{1:4}$. Each covariance matrix, which is the inverse of the inverse, also becomes block-diagonal. Among the three models that have a single edge, model (b) gives the maximum log-likelihood and is regarded to be better than models (c) and (d). In addition, model (b) is regarded as preferable to model (a) in terms of the AIC or the BIC. The other two models, models (c) and (d), have large AIC and BIC values compared with model (a), which indicates that the added edge in these models is unnecessary.

Models (e) through (g) have two edges. Variables 3 and 2 are conditionally independent, given variable 1 in model (e), and similar conditions are assumed for models (f) and (g). Model (f) is described in the first half of this subsection. Among the two-edge models, model (f) gives the maximum log-likelihood, but its AIC and BIC values are larger than those of model (b), which suggests that model (f) is not better than model (b), which has only one edge. Despite having two edges, model (g) is regarded as being worse than the no-edge model, model (a). Comparing the AIC and BIC values among the various models, model (b) is selected as the most preferable covariance that satisfies the regular condition.

5 Implementation

Since vector \mathbf{x} corresponds to the state vector or the observation vector, the vector has a large dimension, namely, n is large. Accordingly, the number of nonzero elements m also becomes large. When no neighbors are assumed, m is equal to n , and m is usually assumed to be larger than n when neighbors are assumed. In this section, we address an implementation technique that is especially useful when n is large.

We assume that the fixed matrices \mathbf{A}_k ($k = 1, \dots, m$) are a set of matrices, each of which has a value of two (instead of a value of unity, as used in Section 4.1) in one of the diagonal elements or has two symmetric off-diagonal elements with values of unity, which form an interacting pair. When we consider the model shown in Figure 2, the regularized inverse Σ^{-1} is expressed as a linear combination of the following five matrices:

$$\begin{aligned} \mathbf{A}_1 &= \begin{pmatrix} 2 & & \\ & & \\ & & \end{pmatrix}, & \mathbf{A}_2 &= \begin{pmatrix} & & \\ & 2 & \\ & & \end{pmatrix}, \\ \mathbf{A}_3 &= \begin{pmatrix} & & \\ & & 2 \\ & & \end{pmatrix}, & \mathbf{A}_4 &= \begin{pmatrix} & & 1 \\ & 1 & \\ 1 & & \end{pmatrix}, \\ \mathbf{A}_5 &= \begin{pmatrix} & & 1 \\ & 1 & \\ 1 & & \end{pmatrix}. \end{aligned} \quad (69)$$

When \mathbf{A}_k has non-zeros in its $(p(k), q(k))$ and $(q(k), p(k))$ positions, \mathbf{A}_k can be represented as

$$\mathbf{A}_k = \mathbf{e}_{p(k)} \mathbf{e}_{q(k)}' + \mathbf{e}_{q(k)} \mathbf{e}_{p(k)}', \quad (70)$$

where \mathbf{e}_i is an n -dimensional unit vector that has unity in its i -th position and zeros elsewhere. Using Eq. (70), we can simplify the cost function given by Eq. (13), the gradient vector given by Eq. (15), and the Hessian matrix








Model	Inverse: $\hat{\Sigma}^{-1}$	Covariance: $\hat{\Sigma}$	$ \hat{\Sigma} $	$\ell(\bar{\mathbf{x}}, \hat{\beta})$	m	AIC	BIC
(a) 	$\begin{pmatrix} 1.131 & & \\ & 1.142 & \\ & & 0.995 \end{pmatrix}$	$\begin{pmatrix} 0.884 & & \\ & 0.876 & \\ & & 1.005 \end{pmatrix}$	0.778	-12.394	3	30.787	28.083
(b) 	$\begin{pmatrix} 5.293 & -4.715 & \\ -4.715 & 5.342 & \\ & & 0.995 \end{pmatrix}$	$\begin{pmatrix} 0.884 & 0.780 & \\ 0.780 & 0.876 & \\ & & 1.005 \end{pmatrix}$	0.167	-10.079	4	28.158	24.553
(c) 	$\begin{pmatrix} 1.132 & & -0.032 \\ & 1.142 & \\ -0.032 & & 0.996 \end{pmatrix}$	$\begin{pmatrix} 0.884 & & 0.028 \\ & 0.876 & \\ 0.028 & & 1.005 \end{pmatrix}$	0.778	-12.392	4	32.785	29.179
(d) 	$\begin{pmatrix} 1.131 & & \\ & 1.500 & -0.684 \\ & -0.684 & 1.307 \end{pmatrix}$	$\begin{pmatrix} 0.884 & & \\ & 0.876 & 0.458 \\ & 0.458 & 1.005 \end{pmatrix}$	0.593	-11.985	4	31.969	28.364
(e) 	$\begin{pmatrix} 5.295 & -4.715 & -0.032 \\ -4.715 & 5.342 & \\ -0.032 & & 0.996 \end{pmatrix}$	$\begin{pmatrix} 0.883 & 0.779 & 0.028 \\ 0.779 & 0.875 & 0.025 \\ 0.028 & 0.025 & 1.005 \end{pmatrix}$	0.166	-10.078	5	30.156	25.649
(f) 	$\begin{pmatrix} 5.293 & -4.715 & \\ -4.715 & 5.700 & -0.684 \\ & -0.684 & 1.307 \end{pmatrix}$	$\begin{pmatrix} 0.884 & 0.780 & 0.408 \\ 0.780 & 0.876 & 0.458 \\ 0.408 & 0.458 & 1.005 \end{pmatrix}$	0.127	-9.670	5	29.340	24.833
(g) 	$\begin{pmatrix} 1.132 & & -0.032 \\ & 1.500 & -0.684 \\ -0.032 & -0.684 & 1.308 \end{pmatrix}$	$\begin{pmatrix} 0.884 & 0.013 & 0.028 \\ 0.013 & 0.876 & 0.458 \\ 0.028 & 0.458 & 1.005 \end{pmatrix}$	0.592	-11.983	5	33.966	29.460

Table I. Summary of the regularized matrices constructed from sample covariance \mathbf{S} given in Eq. (61): The columns show the assumed graphical model, the estimated inverse ($\hat{\Sigma}^{-1}$), the estimated covariance ($\hat{\Sigma}$), the determinant of the covariance ($|\hat{\Sigma}|$), the maximum log-likelihood, the number of parameters (m), the AIC, and the BIC. Both the AIC and the BIC are minimized by model (b).

given by Eq. (16) as

$$f(\beta) = 2 \sum_{k=1}^m \beta_k (\mathbf{S})_{p(k)q(k)} - \log |\Sigma^{-1}| \quad (71)$$

$$\frac{\partial f}{\partial \beta_i} = 2 (\mathbf{S})_{p(i)q(i)} - 2 (\Sigma)_{p(i)q(i)}, \quad (72)$$

$$\frac{\partial^2 f}{\partial \beta_i \partial \beta_j} = 2 (\Sigma)_{p(i)p(j)} (\Sigma)_{q(i)q(j)} + 2 (\Sigma)_{p(i)q(j)} (\Sigma)_{q(i)p(j)}, \quad (73)$$

respectively. Their derivations are given in Appendix C [Eqs. (124), (129), and (136)]. These three equations indicate that we can avoid the matrix multiplications that appear in Eqs. (13), (15), and (16).

5.1 Algorithm

We can minimize the cost function given by Eq. (71) with Newton's algorithm.

Step 1: Select an initial value $\beta^{(0)} \in \mathbb{R}^m$ and tolerance $\varepsilon > 0$. Set an iteration counter $r = 0$.

Step 2: Compute the Newton direction $\Delta\beta^{(r)}$ that satisfies the linear equations

$$[\nabla^2 f(\beta^{(r)})] \Delta\beta^{(r)} = -\nabla f(\beta^{(r)}). \quad (74)$$

Step 3: Quit if $\delta^2(\beta^{(r)}) \leq \varepsilon$, where

$$\delta^2(\beta^{(r)}) = -\nabla f(\beta^{(r)})' \Delta\beta^{(r)}. \quad (75)$$

Step 4: Update $\beta^{(r)}$ to $\beta^{(r+1)}$ according to

$$\beta^{(r+1)} = \beta^{(r)} + 2^{-t} \Delta\beta^{(r)} \quad (76)$$

with minimum $t \in \{0, 1, \dots\}$ that satisfies

$$f(\beta^{(r)} + 2^{-t} \Delta\beta^{(r)}) < f(\beta^{(r)}). \quad (77)$$

Step 5: Set $r = r + 1$ and go to Step 2.

5.2 Remarks

When sample variances differ significantly among variables, we may regularize the sample correlation matrix \mathbf{R} instead of the sample covariance matrix \mathbf{S} for numerical stability. Here, $(\mathbf{R})_{ij}$ is defined as $(\mathbf{S})_{ij} / \sqrt{(\mathbf{S})_{ii} (\mathbf{S})_{jj}}$.

At Step 1, the initial value $\beta^{(0)}$ should be selected such that $\Sigma^{-1}(\beta^{(0)})$ is positive definite. This is a necessary condition under which the right-hand side of Eq. (77), $f(\beta^{(0)})$, is well defined. A convenient way to maintain

the positive definiteness is to assume a diagonal matrix. One example is to select $\beta^{(0)} = (\beta_1^{(0)} \dots \beta_m^{(0)})'$ such that

$$\sum_{k=1}^m \mathbf{A}_k \beta_k^{(0)} = \text{diag} \mathbf{S}^{-I}, \quad (78)$$

where \mathbf{S}^{-I} denotes the generalized inverse of sample covariance matrix \mathbf{S} . Another example, which requires lower computational cost, is to select $\beta^{(0)}$ such that

$$\sum_{k=1}^m \mathbf{A}_k \beta_k^{(0)} = \frac{n}{\text{tr} \mathbf{S}} \mathbf{I}_n, \quad (79)$$

where \mathbf{I}_n is the unit matrix of order n .

At Step 2, the gradient vector $\nabla f(\beta^{(r)})$ and the Hessian matrix $\nabla^2 f(\beta^{(r)})$ are computed with Eqs. (72) and (73), respectively. These two equations require Σ , that is, the inverse of $\Sigma^{-1}(\beta^{(r)})$. Since the latter is a sparse matrix, the computation may be tractable even when n is large. We may use a software packages for sparse matrices. In addition, we may use the conjugate-gradient method to obtain each column of Σ .

Equation (74) is a linear equation of order m . Since m is usually larger than n and the Hessian matrix given by Eq. (73) is a dense matrix, solving Eq. (74) requires the most expensive computational cost. When m is large, the Hessian matrix needs to be distributed over multiple memories. Since the Hessian matrix is positive definite, we may use a solver, such as PDPOSV in ScaLAPACK, which uses the Cholesky decomposition to factor a distributed positive definite matrix, for parallel processing.

In Step 3, $\delta(\beta^{(r)})$ is the Newton decrement, originally defined as (Boyd and Vandenberghe, 2004, Section 9.5.1)

$$\delta(\beta^{(r)}) = \sqrt{\nabla f(\beta^{(r)})' \nabla^2 f(\beta^{(r)})^{-1} \nabla f(\beta^{(r)})}. \quad (80)$$

The stopping criterion uses $\delta^2(\beta^{(r)})$, because $\delta^2(\beta^{(r)})$ is an estimate of $f(\beta^{(r)}) - \min_{\beta} f(\beta)$, and $-\delta^2(\beta^{(r)})$ is interpreted as the directional derivative of $f(\beta)$ at $\beta^{(r)}$ in the Newton direction (Boyd and Vandenberghe, 2004, Section 9.5.1). We set the tolerance $\varepsilon = 10^{-10}$ in the following applications in Section 6.

At Step 4, the Newton direction $\Delta\beta^{(r)}$ is scaled by 2^{-t} . Using the scale parameter 2^{-t} , we conduct a line search that roughly finds a minimum along the Newton direction. The scale parameter does not exist in the pure form of Newton's method, but the parameter is introduced to guard against the possibility that the cost function might increase due to non-quadratic terms in the cost function.

Evaluating Eq. (77) requires $\log |\Sigma^{-1}|$. Again, since Σ^{-1} is a sparse matrix, this term can be computed without a huge computational cost. Note that as long as $\beta^{(r)}$ is

updated while satisfying Eq. (77), the positive definiteness of $\Sigma(\beta^{(r)})$ is guaranteed. If at least one of the eigenvalues takes a value of zero or becomes negative, the second term of Eq. (71) cannot be defined and consequently the condition given by Eq. (77) does not hold.

6 Application

We apply the proposed regularization method to sample covariance matrices constructed from the TOPEX/POSEIDON (T/P) altimetry observation. The sample covariance is calculated from the SSH anomalies, the trend component of which was subtracted (Ueno *et al.*, 2007). The total number of the time steps is $N_e = 364$ (T/P cycles from September 23, 1992 to August 11, 2002).

In the following two subsections, we regularize the sample covariance matrices in the equatorial Pacific Ocean and in the global ocean, respectively. The dimension of the variable is $n = 503$, and $n = 8585$ in the two applications. In any case, the $n \times n$ matrix is rank deficient (its rank is less than N_e) and is therefore singular. In the first application, we show the overall procedures for regularization and examine the obtained matrices, while we aim to demonstrate the feasibility of the regularization process for a large-dimensional covariance matrix in the second application.

6.1 Application 1: Equatorial Pacific Ocean

6.1.1 Sample covariance

We regularize a sample covariance matrix obtained for data points of every $L_x = 8^\circ$ in the zonal direction and $L_y = 2^\circ$ in the meridional direction in the equatorial Pacific Ocean. The number of data points becomes $n = 503$. Figure 4a shows the sample covariance matrix \mathbf{S} of the detrend observations. Diagonal elements are dominated, but non-diagonal elements also appear to have meaningful values. Elements in $[278, 364] \times [278, 364]$, denoted by a dashed square in Figure 4a, are enlarged in Figure 4b. We can identify a pattern of non-diagonal elements, for example, elements of positive-negative pairs around (293, 323) and (322, 353), and positive elements around (290, 350). Diagonal elements, i.e., the variance, are shown in Figure 4c. The meshes outlined in white, those in $148^\circ\text{--}124^\circ\text{W}$ and $28^\circ\text{S--}30^\circ\text{N}$, correspond to the enlarged area shown in Figure 4b. Figure 4d shows the elements in the 323-rd row, which correspond to $(136^\circ\text{W}, 5^\circ\text{N})$, and are outlined in white. While positive covariances are identified in the meridional direction, strong negative covariances appear at neighboring observation points aligned in the zonal direction. The positive-negative pairs around (293, 323) in Figure 4b correspond to positive covariances around $(144^\circ\text{W}, 1^\circ\text{S})$ and negative covariances around $(144^\circ\text{W}, 7^\circ\text{N})$.

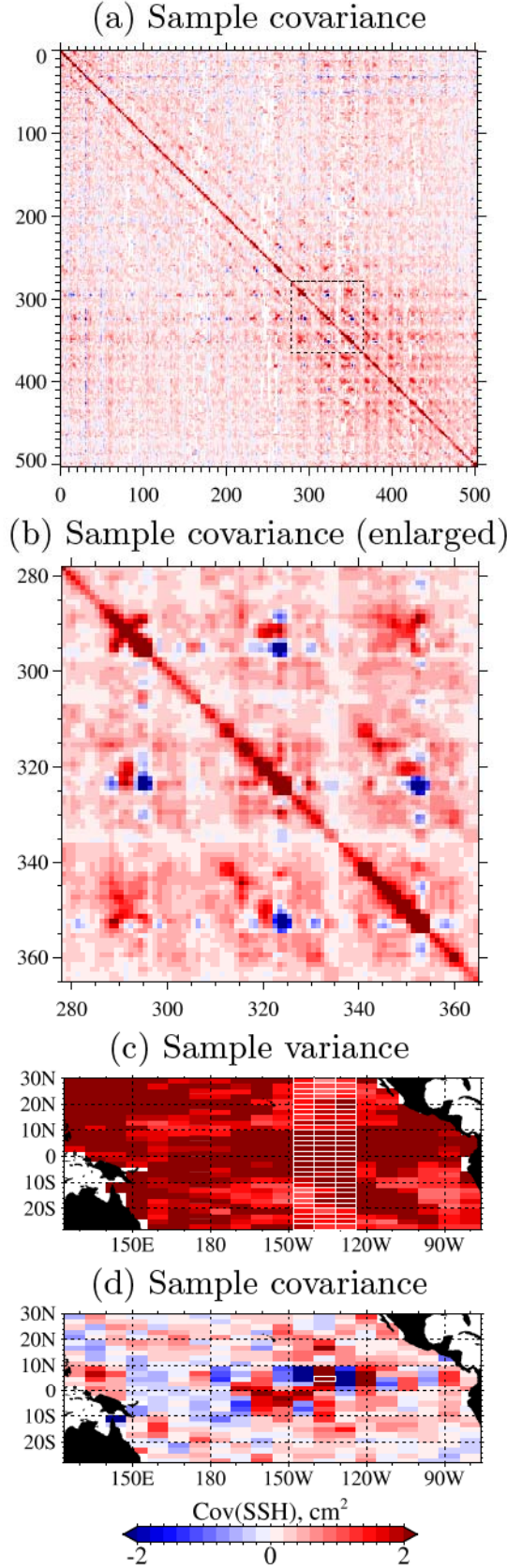


Figure 4. Sample covariance calculated from detrend SSH observations: (a) sample covariance matrix, (b) elements in [278, 364], (c) variance (diagonal elements of the covariance matrix), and (d) elements in the 323-rd row (corresponding to (136°W, 5°N), outlined in white). The enlarged area in panel (b) corresponds to the dashed square in panel (a) and the meshes outlined in white in panel (c).

6.1.2 Regularization in inverse space

We regularize the sample covariance matrix shown in Figure 4. We apply eight covariance selection models shown in Figure 3, and estimate their parameters by the maximum likelihood with Newton's method. Figure 5 shows estimated inverse covariances $\hat{\Sigma}^{-1}$ and corresponding covariances $\hat{\Sigma}$. The squares in the inverse represent prescribed non-zero elements. When no neighbors are allowed (Figure 5a), the inverse becomes diagonal, and the corresponding covariance of course becomes diagonal. When neighbors are assumed, corresponding covariances have finite values in all elements. When four neighbors are assumed (Figure 5b), although most of the elements of the inverse are zero, the corresponding covariance has a pattern similar to that of the sample covariance (Figure 4b), which includes the negative elements around (325, 350), for example. The pattern of covariance elements is similar to that in the case of eight neighbors (Figure 5c). In the case of 12 neighbors (Figure 5d), the corresponding covariance becomes less faint, and larger values exist only around the elements of nonzero elements of the inverse. When 20-, 24-, 28-, and 36-neighbor models are assumed (Figures 5e, 5f, 5g, and 5h), the covariance pattern appears to be faint in accordance with the increasing number of nonzero elements.

6.1.3 Preferable number of neighbors

Among the covariance matrices shown in Figure 5, we select a preferable covariance with the AIC or BIC. Figure 6 shows AIC and BIC variations as a function of the number of neighbors, and their actual values are summarized in Table II. Both the AIC and BIC values decrease significantly (approximately 10^5) even when the minimum number of neighbors, namely four, is assumed. This suggests that non-diagonal matrices are far preferable compared to the simple diagonal matrix. With the eight-neighbor model, while the AIC continues to decrease, the BIC increases, and neighbors located in oblique directions are considered to be ineffective. We may stop the application of the covariance selection models here and select the four-neighbor model as an appropriate model.

When we apply the 12-neighbor model, both the AIC and the BIC have a minimum. For the BIC, the value for the 12-neighbor model is smaller than that for the four-neighbor model and is smaller than those for with models having more neighbors. The 12-neighbor model is found to have the minimum BIC value. The AIC profile is minimum, which is found to be a global minimum (Table II), for the 28-neighbor model. From the examination of the AIC and BIC profiles, one of the following options may be selected: (1) the four-neighbor model, because the BIC has the first minimum, (2) the 12-neighbor model, because the AIC has the first minimum and/or the BIC has the global minimum, (3) the 28-neighbor model, because the AIC has the global minimum. Here, we will select the covariance using the 12-neighbor model (Figure 5d) as the preferable model among the eight trial models. Figure 7 shows the selected covariance. Band structures from

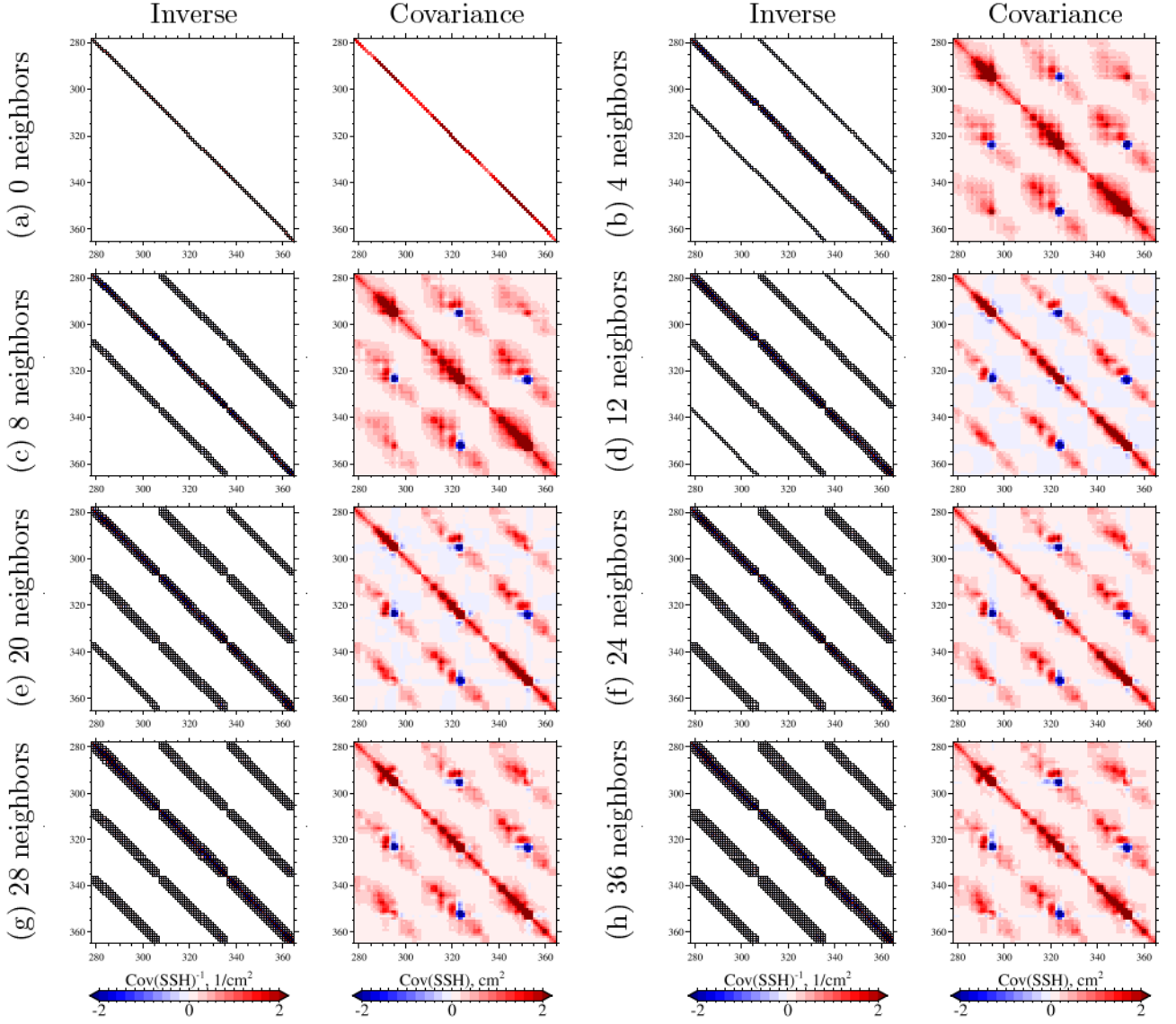


Figure 5. Estimated inverse covariances and corresponding covariances obtained with (a) zero-, (b) four-, (c) eight-, (d) 12-, (e) 20-, (f) 24-, (g) 28-, and (h) 36-neighbor models. Only elements in $[278, 364]$ are shown. The squares shown in the inverse matrices represent prescribed non-zero elements.

the upper-left to the lower-right appear to be emphasized, while elements between the bands are reduced to small values, which means that correlations for distant points are reduced, as shown in Figure 7d. The estimated variances shown in Figure 7c are identical to those based on sample statistics (Figure 4c).

6.2 Application 2: Global ocean

6.2.1 Sample covariance and regularized covariances

We apply the covariance selection models to a sample covariance of the detrend SSH observation in the global ocean. We use all of the gridded data of spatial resolution of $L_x = 2^\circ$ in the zonal direction and $L_y = 2^\circ$ in the meridional direction. The number of data points is equal to $n = 8,585$. For simplicity, we neglect the neighboring relation between variables across a longitude of 0° . This

assumption would be acceptable because we aim simply to demonstrate the applicability of the proposed algorithm (Section 5) to a large-dimensional matrix.

Figure 8 compares the sample covariance S and regularized covariances $\hat{\Sigma}$ that are obtained by assuming different numbers of neighbors. The variance elements of $\hat{\Sigma}$, shown in the left-hand panels of Figures 8b through 8i, appear to be identical to those of S (Figure 8a). This indicates that β converged to the optimum value in each covariance model.

The right-hand side of Figure 8a shows the sample covariance with respect to the variable at $(180^\circ, 0^\circ)$. The covariance elements display correlated variables that are confined in 162°E - 162°W and 6°N - 6°S , and anti-correlated variables in 105° - 136°W along the equator. In addition, the stripe structures along the satellite orbit are identified in off-equatorial to mid-latitude regions in the

Number of neighbors	$\log \hat{\Sigma} $	$\ell(\bar{\mathbf{x}}, \hat{\beta})$	m	AIC	BIC
0	418.4	-295336.7	503	591679.3	593574.8
4	54.1	-237047.2	1457	477008.4	482498.9
8	43.6	-235373.2	2375	475496.5	484446.2
12	9.2	-229860.5	3278	466277.0	478629.5
20	-1.7	-228122.3	5028	466300.5	485247.7
24	-6.7	-227324.6	5863	466375.2	488468.9
28	-15.0	-225985.6	6716	465403.2	490711.2
36	-22.8	-224741.8	8374	466231.5	497787.4

Table II. Summary of covariances estimated with the covariance selection models: the number of neighbors, $\log |\hat{\Sigma}|$, the maximum likelihood ($\ell(\bar{\mathbf{x}}, \hat{\beta})$), the number of parameters ($m = \dim \beta$), and the AIC and BIC values for models in which different numbers of neighbors are assumed. A local minimum of the AIC is obtained for the 12-neighbor model, and the global minimum of the AIC is obtained for the 28-neighbor model. For the BIC, a local minimum is obtained for the 4-neighbor model, and the global minimum is obtained for the 12-neighbor model.

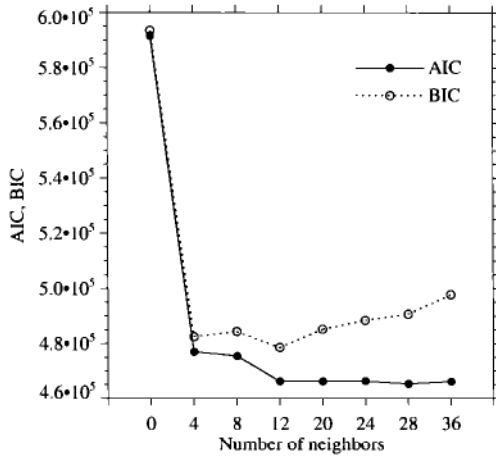


Figure 6. AIC and BIC values as a function of the number of neighbors in an application for the equatorial Pacific Ocean.

Pacific and Indian Oceans. A spurious covariate structure that arises due to the small ensemble size appears in the mid-latitude region in the Atlantic Ocean.

The estimated covariance $\hat{\Sigma}$ with zero neighbors is a diagonal matrix that is identical to $\text{diag } S$, and no covariate variables are estimated, as shown in Figure 8b. With the four-neighbor model, correlated variables are estimated inside a nearly circular region of a radius of approximately 20° (Figure 8c). A similar circular region is also identified with the eight-neighbor model, but its radius decreases to 15° (Figure 8d). When 12 neighbors are assumed, the region for correlated variables becomes an elliptic form that is spread zonally from 160°E to 148°W and is confined meridionally within 6°N to 6°S (Figure 8e). The 20-neighbor model (Figure 8f) also estimates an elliptic region that is spread zonally, but that is slightly more confined in the meridional direction. Such elliptic regions also appear with the 24- to 36-neighbor models (Figures 8g through 8i), and the appearance of these regions remains unchanged. Regardless of the number of neighbors, the regularized covariances shown above do

not reproduce the anti-correlated variables in the central-to-eastern equatorial Pacific Ocean, the stripe structure in the Pacific and Indian Oceans, or the presumed spurious structure in the mid-latitude Atlantic Ocean.

6.2.2 Preferable number of neighbors

Figure 9 shows the AIC and BIC values as functions of the number of neighbors, and Table III summarizes the results. Both the AIC and BIC values increase drastically when four neighbors are assumed, compared to the case in which it is assumed that there are no neighbors. When we assume eight neighbors, the AIC and the BIC decrease, but these values are roughly comparable with the four-neighbor model. As shown in Figures 8c and 8d, both the four- and eight-neighbor models estimated circular regions of correlated variables. When 12 neighbors are assumed, the AIC and BIC values decrease significantly compared with the eight-neighbor model, that is, the quantity $\text{AIC}(8) - \text{AIC}(12)$ ($\text{BIC}(8) - \text{BIC}(12)$) is larger than the quantity $\text{AIC}(4) - \text{AIC}(8)$ ($\text{BIC}(4) - \text{BIC}(8)$). In addition, when 12 or more neighbors are assumed, the AIC and the BIC, respectively, take comparable values independent of the number of neighbors. This implies that the 12- to 36-neighbor models commonly estimate the elliptic regions of correlated variables. Among the models with 12 or more neighbors, the AIC continues to decrease as the number of neighbors increases, which suggests that the 36-neighbor model is the most appropriate model. The BIC profile, however, shows the first minimum on the 20-neighbor model, and the second (and global) minimum on the 28-neighbor model. With the BIC, we can select the 20-neighbor or 28-neighbor model as the preferable model.

7 Discussion

We have developed a method for the regularization of sample covariance matrices that is carried out in inverse space. Here, we first address (1) a difficulty in conducting such a regularization in physical space, (2) a filtering effect of the regularization on spurious correlation, (3) the

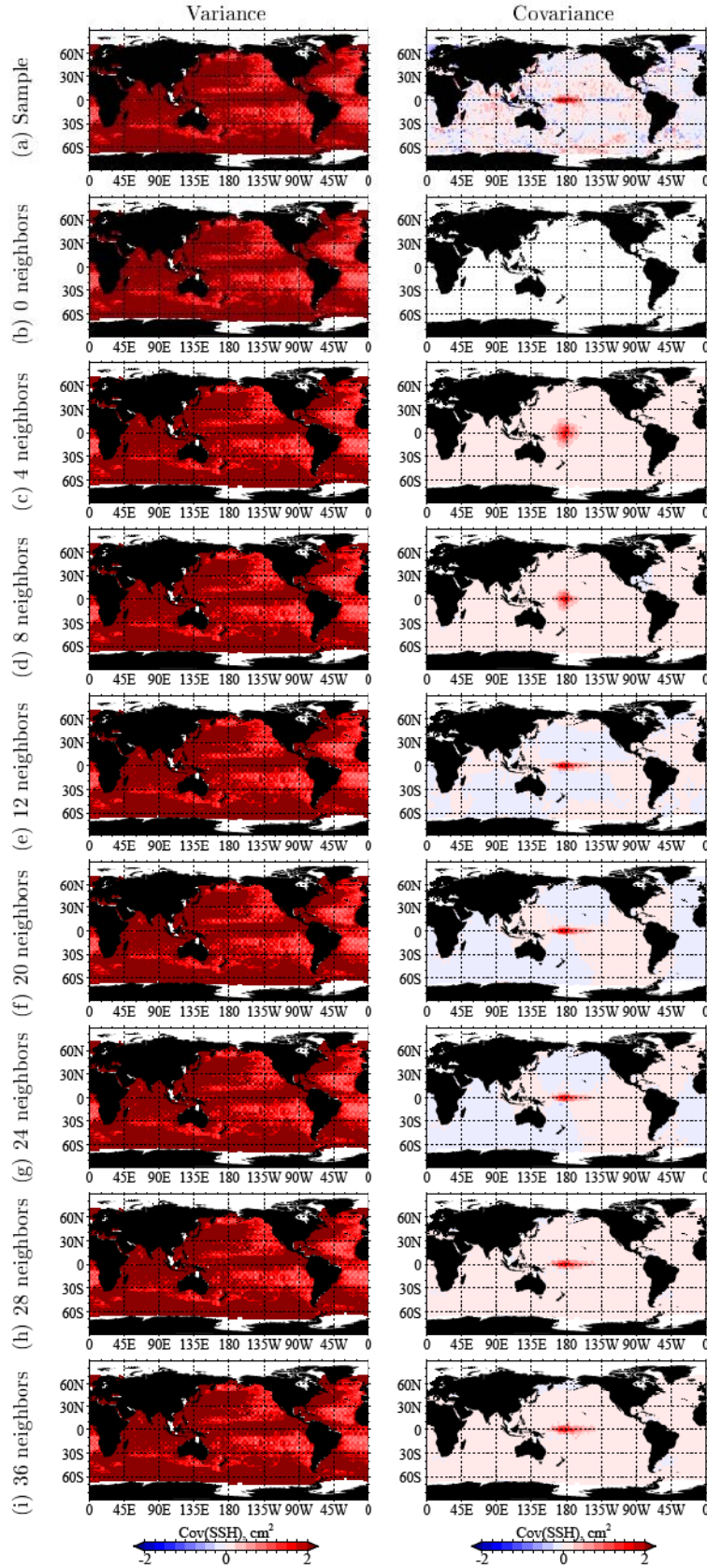


Figure 8. Variance and covariance elements of detrend SSH observation at $2^\circ \times 2^\circ$ grid points in the global ocean. The covariance elements are those with the variable at $(180^\circ, 0^\circ)$. Panel (a) shows the sample covariance S . Panels (b) through (j) show the regularized covariance $\hat{\Sigma}$ obtained by assuming 0, 4, 8, 12, 20, 24, 28, and 36 neighboring variables, respectively.

difference from a matrix approximation with the singular value decomposition (SVD), and (4) a computational cost

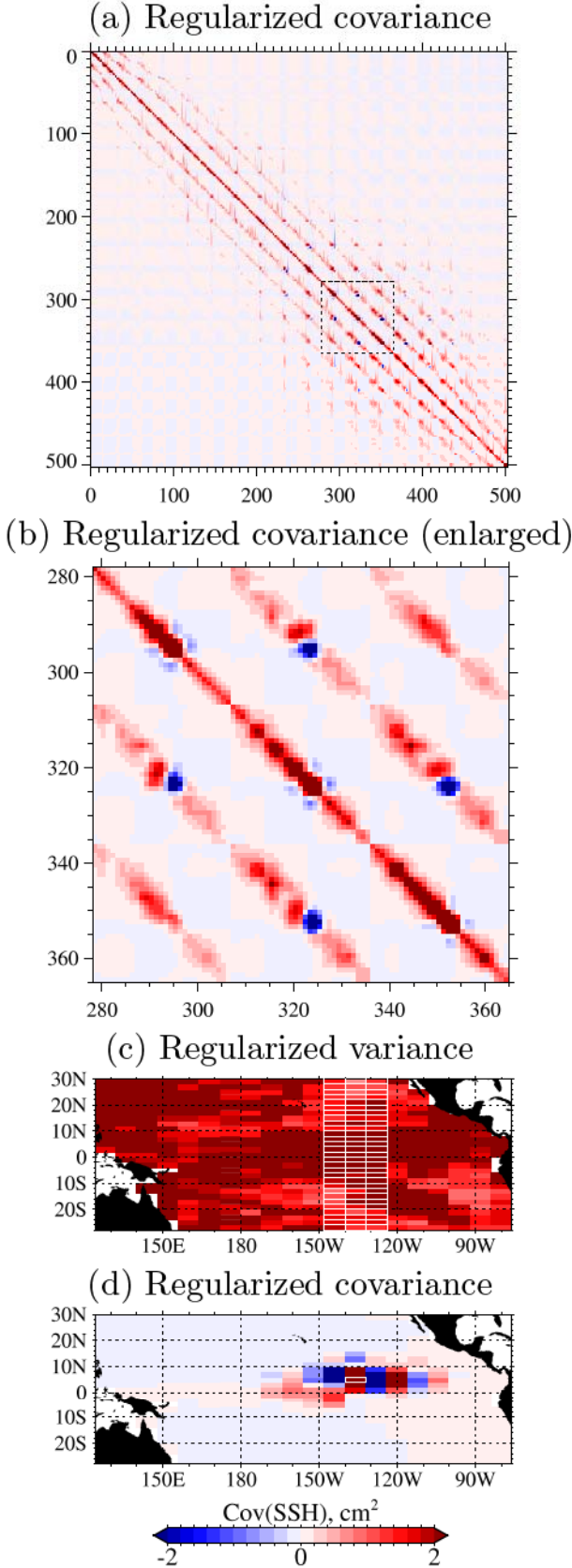


Figure 7. Regularized covariance estimated with the sample covariance (Figure 4) and the 12-neighbor model (Figure 3d). The format is the same as that in Figure 4.

reduction in data assimilation brought by the inverse space

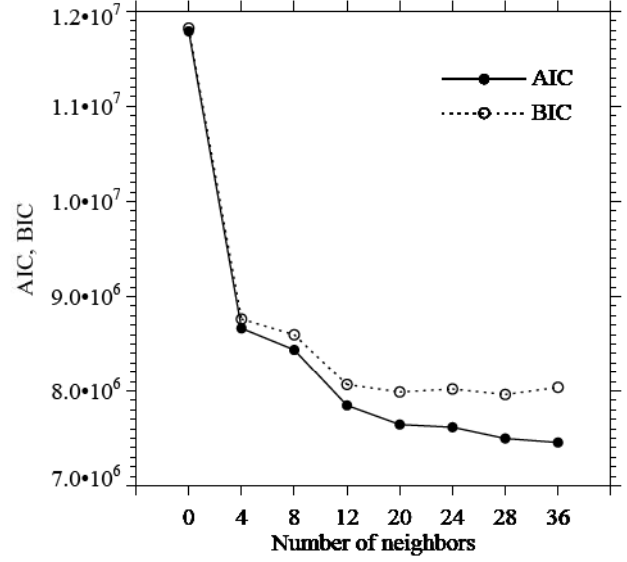


Figure 9. AIC and BIC values as functions of the number of neighbors in an application for the global ocean.

modeling. We then discuss the (5) pre-conditioning of variables using the regularized covariance, (6) a problem that may occur with a very small number of samples, and (7) a perspective for larger problems.

7.1 Modeling of independence and conditional independence

The most characteristic feature of the regularization in inverse space is the explicit awareness that a covariance matrix is a parameter that specifies a Gaussian distribution. If variables obey a Gaussian distribution, conditional independence between a pair of variables corresponds to a zero element of the inverse of the covariance matrix (e.g., [Lauritzen, 1996](#), Proposition 5.2). Modeling of conditional independence is easy to perform compared to the modeling of independence, because only the local relationship between a pair of variables need be taken into account. For example, the relationship between the three variables shown in Figure 2, which indicates that variables 1 and 3 are conditionally independent given variable 2, is difficult to model with usual covariance. Since variables 1 and 3 are not independent (i.e., dependent on each other through variable 2), we cannot impose an explicit constraint (for example, forcing some elements to be zeros) on the covariance matrix.

In addition, the straightforward concept of specifying independence may yield an invalid covariance unless the modeled matrix is diagonal or block diagonal. As an example of a covariance that is neither diagonal nor block diagonal, we assume independence between variables 1 and 3. The estimated covariance then becomes

$$\hat{\Sigma} = \begin{pmatrix} s_{11} & s_{12} & 0 \\ s_{12} & s_{22} & s_{23} \\ 0 & s_{23} & s_{33} \end{pmatrix}. \quad (81)$$

By the definition of the Gaussian distribution, $\hat{\Sigma}$ should be positive definite. This condition is equivalent to the

existence of the Cholesky decomposition of $\hat{\Sigma}$, but it is not always satisfied and requires that

$$\frac{s_{12}^2}{s_{11}s_{22}} + \frac{s_{23}^2}{s_{22}s_{33}} < 1. \quad (82)$$

Apparently, this condition does not always hold, depending on S .

7.2 Filtering of spurious correlation

As noted in the Introduction, compactly supported correlation functions have been used to remove spurious correlations due to the limited number of samples (ensemble members). Our regularization in inverse space can have similar effects that filter out the spurious distant correlations.

The regularized covariance given by Eq. (46) corresponds to a correlation matrix

$$\hat{\mathbf{P}} = \begin{pmatrix} 1 & \frac{s_{12}}{\sqrt{s_{11}s_{22}}} & \frac{s_{12}}{\sqrt{s_{11}s_{22}}} \frac{s_{23}}{\sqrt{s_{22}s_{33}}} \\ \frac{s_{12}}{\sqrt{s_{11}s_{22}}} & 1 & \frac{s_{23}}{\sqrt{s_{22}s_{33}}} \\ \frac{s_{12}}{\sqrt{s_{11}s_{22}}} \frac{s_{23}}{\sqrt{s_{22}s_{33}}} & \frac{s_{23}}{\sqrt{s_{22}s_{33}}} & 1 \end{pmatrix} \quad (83)$$

The correlation matrix $\hat{\mathbf{P}}$ indicates that the correlation between variables 1 and 3 has a smaller magnitude than that between variables 1 and 2, because

$$\left| \frac{s_{12}}{\sqrt{s_{11}s_{22}}} \frac{s_{23}}{\sqrt{s_{22}s_{33}}} \right| = \left| \frac{s_{12}}{\sqrt{s_{11}s_{22}}} \right| \left| \frac{s_{23}}{\sqrt{s_{22}s_{33}}} \right| \quad (84)$$

$$\leq \left| \frac{s_{12}}{\sqrt{s_{11}s_{22}}} \right|. \quad (85)$$

This means that a monotonically decreasing correlation will be obtained when the direct relationship between variables is assumed according to the distance between the grid points. In addition, as described in Section 3, regularized correlations take values identical to the sample correlations for nearby grid points, where non-zero elements are assumed in the inverse matrix. These two properties are as expected for a filtering function that removes spurious correlations for distant grid points while maintaining the original correlation for nearby grid points. The properties are not always expected in a sample correlation. The correlation between variables 1 and 3, $s_{13}/\sqrt{s_{11}s_{33}}$, which may be contaminated by spurious correlation, can be larger than that between variables 1 and 2, $s_{12}/\sqrt{s_{11}s_{22}}$.

7.3 Comparison with an SVD-approximated matrix

The covariance selection is considered to be an approach for approximating a sample covariance matrix. For a matrix approximation, the singular value decomposition (SVD) technique has been widely applied. With the SVD, we decompose a matrix into singular vectors and singular values, neglect small singular values by forcing them

to be zero, and again multiply the singular vectors and modified singular values to form an approximated matrix. The approximated matrix is based on the sample covariance alone and does not take the distance between the grid points into account. The SVD approximation is, therefore, not expected to work as a procedure that preferentially removes spurious correlations for distant points. In fact, Figure 10 shows an SVD-approximated sample covariance, using 23 leading singular values that are larger than 10^{-2} times the sum of all of the singular values. Correlations for distant points appear to remain (Figure 10d), while the variances are reduced (Figure 10c) compared to the original sample variances (Figure 4c).

The dependence of the number of singular values used in the SVD approximation is shown in Figure 11. Variances increase gradually with the number of singular values used, and account for 49% and 94% of the original sample variance (Figure 11h) when 23 and 155 leading terms are used (Figures 11f and 11g). On the other hand, covariance elements for the variable at (136°W, 5°N) appear to be well represented with at least two leading singular values (Figures 11b). Even with the two leading terms, correlations for distant points are also reproduced. This indicates that the SVD approximation does not work as a filter of the long-distance correlations.

In addition, an inverse matrix also approximated with the SVD (referred to as the generalized inverse). The approximated inverse generally has non-zero values in all elements. This means that the SVD approximation is not consistent with the assumption of conditional independence, which is adopted in the covariance selection model.

7.4 Reduction of the number of parameters

A covariance matrix has elements of the square of the number of variables, amounting to 10^{14} in data assimilation. One of the purposes of covariance regularization is to reduce the number of parameters that specify the covariance matrix. Among the other regularization methods reviewed in Section 1, the proposed regularization method in inverse space can significantly reduce the number of parameters for the matrix specification. When the four-neighbor model is assumed, the required number of parameters is approximately four times the number of variables, which is much smaller than the square of the number of variables.

The reduced number of parameters in inverse space also reduced the computational costs in evaluating cost functions in variational data assimilation. Since the covariances of background error and observation noise are evaluated in their inverted forms in cost functions, many zeros in the modeled inverse matrices significantly reduce the number of multiplications.

7.5 Pre-conditioning with regularized covariances in inverse space

The regularized covariances in inverse space can be used for pre-conditioning (Courtier *et al.*, 1994). In 4D-Var with a background error covariance matrix \mathbf{B} , the

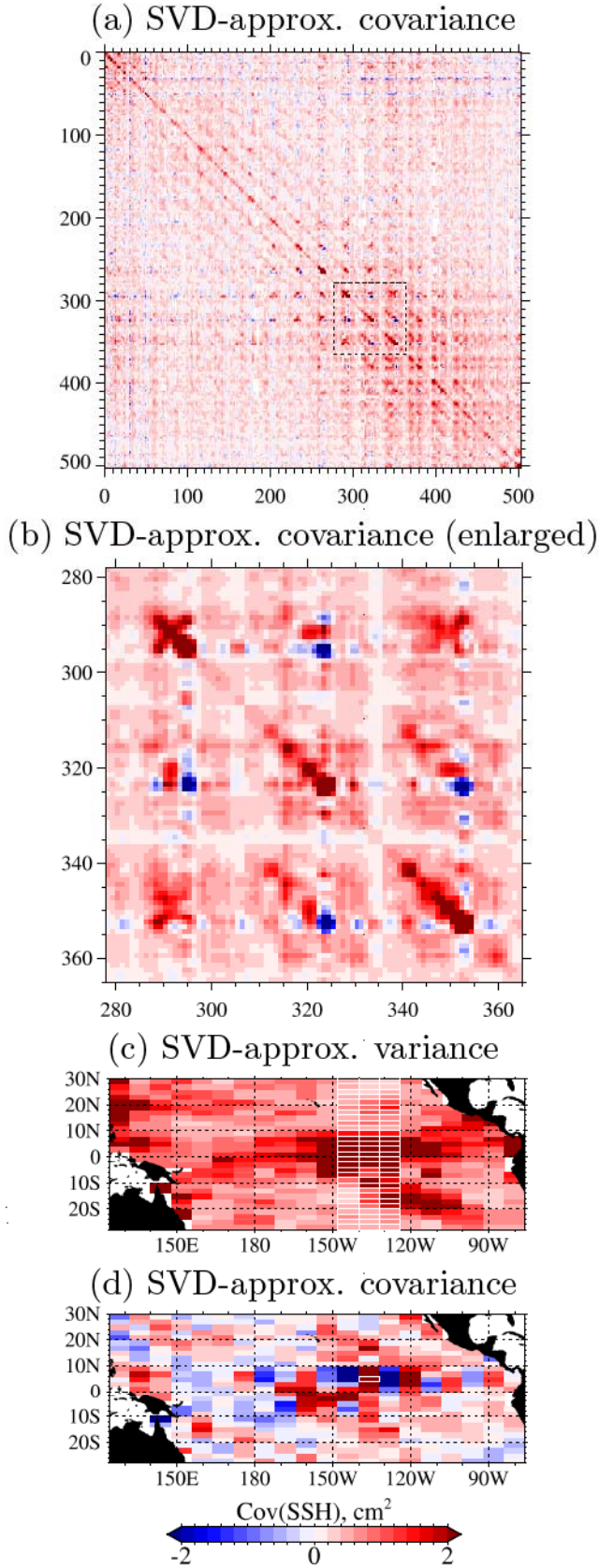


Figure 10. The SVD-approximated sample covariance using 23 leading singular values that are larger than 10^{-2} times the sum of all of the singular values. The format is the same as that in Figure 4.

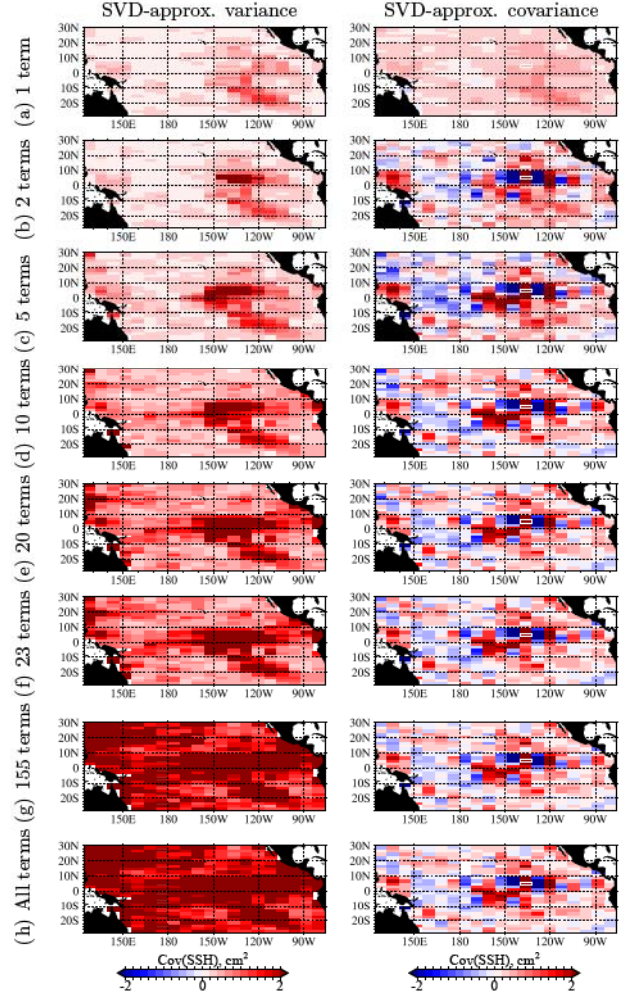


Figure 11. SVD-approximated sample covariances using (a) 1, (b) 2, (c) 5, (d) 10, (e) 20, (f) 23, (g) 155, and (h) all leading singular values. The left-hand panels show the variance (diagonal elements of the covariance matrices), and the right-hand panels show the covariance elements for the variable at $(136^\circ\text{W}, 5^\circ\text{N})$, which is outlined in white. The 23 and 155 leading terms (panels (f) and (g)) correspond to the singular values that are larger than 10^{-2} and 10^{-3} times the sum of all of the singular values, respectively. The covariance using all singular values (panel (h)) is identical to the sample covariance.

square-root of the matrix $B^{1/2}$ is usually selected as a pre-conditioner matrix. Assuming \mathbf{x} and \mathbf{x}_b to be a state vector and its background estimate (prescribed mean), respectively, the background cost is represented as $(\mathbf{x} - \mathbf{x}_b)^T B^{-1} (\mathbf{x} - \mathbf{x}_b) / 2$. A new variable \mathbf{u} defined as $\mathbf{u} = B^{-1/2} (\mathbf{x} - \mathbf{x}_b)$ transforms the background cost to $\mathbf{u}^T \mathbf{u} / 2$. That is, the transformed variable \mathbf{u} eliminates an explicit evaluation of B^{-1} in the background cost and decreases the condition number of the Hessian matrix of the cost function, which increases the convergence rate of an iterative process for minimization. In the observation cost, on the other hand, the state vector \mathbf{x} must be evaluated with \mathbf{u} , as $\mathbf{x} = \mathbf{x}_b + B^{1/2} \mathbf{u}$.

With a regularized background inverse covariance $\hat{\Sigma}^{-1}$, which we here assume to be an approximation of B^{-1} , such a pre-conditioning process, can be applied. After the preconditioning, the inverse covariance $\hat{\Sigma}^{-1}$

does not appear in the cost function, while $\hat{\Sigma}^{1/2}\mathbf{u}$ is required. The latter term can be obtained through the following two steps: (1) Factorize $\hat{\Sigma}^{-1}$ by the Cholesky decomposition to obtain $\hat{\Sigma}^{-1/2}$, and (2) Solve the equation $\mathbf{u} = \hat{\Sigma}^{-1/2}\mathbf{v}$ for \mathbf{v} , and the obtained \mathbf{v} is $\hat{\Sigma}^{1/2}\mathbf{u}$, which is required in the observation cost. The Cholesky decomposition will be tractable because $\hat{\Sigma}$ is sparse. The linear equation $\mathbf{u} = \hat{\Sigma}^{-1/2}\mathbf{v}$ can be solved without a huge cost because $\hat{\Sigma}^{-1/2}$ is sparse and triangular.

The pre-conditioning by $\hat{\Sigma}^{1/2}$ takes over the properties of $\hat{\Sigma}$, so that it can eliminate spurious correlations that may be present in the original covariance \mathbf{B} .

7.6 Estimated variances and covariances for neighboring variables

An estimated covariance matrix $\hat{\Sigma}$ has values that are identical to the original sample covariance \mathbf{S} at positions where the non-zero elements were assumed in Σ . This means that the estimated variances and covariances for neighboring variables maintain the sample statistics. Sample variances and covariances for variables located at nearby points are expected to take values close to the true values (e.g., Bishop and Hodyss, 2007, Figure 1(a)), even when a finite number of samples are used. In the case of a very limited number of samples, however, the deviation from the true value may become large. We may take the spatial average (Raynaud *et al.*, 2008) of \mathbf{S} or apply a spatiotemporal filter (Keppenne *et al.*, 2008) to \mathbf{S} before applying the covariance selection model proposed in the present paper.

7.7 Perspective for large problems

In Section 6.2, we have presented an application of an $n = 8,585$ -dimensional covariance matrix that has $m = 150,381$ parameters. The number is, however, not large enough to directly deal with state-of-the-art dynamic models that may have $n = O(10^7)$ variables. Although there is a gap on the order of 10^3 , the upper limit of the proposed method is expected to increase up to 10^6 by tuning the implementing procedure. We have succeeded in applying a four-neighbor model to global SSH observation at each $1^\circ \times 1^\circ$ grid point, which amounts to $n = 34,300$ grid points and $m = 101,310$ nonzero parameters.

An alternative method of dealing with large problems is to divide the entire domain into multiple regions. A block diagonal structure is then assumed in the covariance matrix Σ , and, therefore, Σ^{-1} , which greatly reduces the computational cost. The regional approach may be justified because effective correlating areas are confined, as shown in the applications (Section 6). The size of an effective correlating area can be estimated from the original sample covariance, \mathbf{S} (e.g., Pannekoucke *et al.*, 2008).

8 Conclusion

We have proposed a method for the regularization of covariance matrices in inverse space with the covariance selection model (the Gaussian graphical model). For

each variable, we prescribed its neighboring variables. The targeted variable is directly related to the neighbors and is conditionally independent of the variables beyond the neighbors. Conditional independence is expressed by specifying zero elements in the inverse covariance matrix. The non-zero elements were estimated numerically using the maximum likelihood with Newton's method. The gradient vector and the Hessian matrix required in Newton's method were derived analytically for the covariance selection model. Appropriate neighboring variables can be selected with information criteria such as the AIC and the BIC. We have presented a technique for implementation that is useful when the dimension of the variable is large. Using the proposed method, the sample covariance shown in Figure 4 is converted to a regularized covariance shown in Figure 7. In addition, we demonstrated the regularization of a $8,585 \times 8,585$ sample covariance having up to 150,381 parameters, which corresponds to $2^\circ \times 2^\circ$ resolution data of the global ocean obtained by TOPEX/POSEIDON altimetry. The regularization gives non-singular covariances and filters out spurious correlations at distant points. In addition, the regularized covariances can be specified by the small number of parameters compared with the square of the dimension of the variables. The regularization method in inverse space will provide a positive contribution in data assimilation, as was the case for the methods formerly developed in physical, spectral, and wavelet space.

Acknowledgement

The Altimeter Ocean Pathfinder TOPEX/POSEIDON SSH anomaly data were provided by the Physical Oceanography Distributed Active Archive Center (PO.DAAC) at the NASA Jet Propulsion Laboratory, Pasadena, California, through their Web site at <http://podaac.jpl.nasa.gov/>. This research was partially supported by the Japan Science Technology Agency for CREST (Core Research for Evolutional Science and Technology) project.

A Derivation of log-likelihood (Eq. 6)

We start with Eq. (5):

$$\ell(\boldsymbol{\mu}, \Sigma) = N_e \left(-\frac{n}{2} \log 2\pi - \frac{1}{2} \log |\Sigma| \right) - \frac{1}{2} \sum_{i=1}^{N_e} (\mathbf{x}_i - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}). \quad (86)$$

The third term multiplied by (-2) can be expanded as

$$\begin{aligned} & \sum_{i=1}^{N_e} (\mathbf{x}_i - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \\ &= \sum_{i=1}^{N_e} (\mathbf{x}_i - \bar{\mathbf{x}})' \Sigma^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) \\ &+ N_e (\bar{\mathbf{x}} - \boldsymbol{\mu})' \Sigma^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}), \end{aligned} \quad (87)$$

where $\bar{\mathbf{x}}$ is the sample mean given by Eq. (1). The first term of Eq. (87) can be expressed as

$$\sum_{i=1}^{N_e} (\mathbf{x}_i - \bar{\mathbf{x}})' \Sigma^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) = N_e \text{tr} \mathbf{S} \Sigma^{-1}, \quad (88)$$

where \mathbf{S} is the sample covariance given by Eq. (2). With Eqs. (87) and (88), the log-likelihood given by Eq. (86) becomes

$$\ell(\boldsymbol{\mu}, \Sigma) = -\frac{N_e}{2} [n \log 2\pi - \log |\Sigma^{-1}| + \text{tr} \mathbf{S} \Sigma^{-1} + (\bar{\mathbf{x}} - \boldsymbol{\mu})' \Sigma^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu})], \quad (89)$$

which we have shown in Eq. (6).

B Derivation of the gradient and the Hessian

We derive the gradient vector given by Eq. (15) and the Hessian matrix given by Eq. (16) of a function given by Eq. (13):

$$f(\beta) = \text{tr} \mathbf{S} \Sigma^{-1} - \log |\Sigma^{-1}|, \quad (90)$$

with respect to $\beta = (\beta_1 \cdots \beta_m)'$. Here, Σ^{-1} is assumed to be parameterized as given in Eq. (9):

$$\Sigma^{-1}(\beta) = \sum_{k=1}^m \mathbf{A}_k \beta_k, \quad (91)$$

where \mathbf{A}_k ($k = 1, \dots, m$) is a fixed matrix (symmetry is not required). From Eq. (91), the first term of Eq. (90) becomes

$$\text{tr} \mathbf{S} \Sigma^{-1} = \sum_{k=1}^m \beta_k \text{tr} \mathbf{S} \mathbf{A}_k. \quad (92)$$

B.1 Gradient vector

Differentiating $f(\beta)$ with respect to β_i ($i = 1, \dots, m$) gives

$$\frac{\partial f}{\partial \beta_i} = \frac{\partial \text{tr} \Sigma^{-1} \mathbf{S}}{\partial \beta_i} - \frac{\partial \log |\Sigma^{-1}|}{\partial \beta_i}. \quad (93)$$

With Eq. (92), the first term becomes

$$\begin{aligned} \frac{\partial \text{tr} \Sigma^{-1} \mathbf{S}}{\partial \beta_i} &= \frac{\partial}{\partial \beta_i} \sum_{k=1}^m \beta_k \text{tr} \mathbf{S} \mathbf{A}_k \\ &= \text{tr} \mathbf{S} \mathbf{A}_i \end{aligned} \quad (94)$$

We expand the second term multiplied by (-1) with the elements of Σ^{-1} as

$$\frac{\partial \log |\Sigma^{-1}|}{\partial \beta_i} = \frac{\partial \log |\Sigma^{-1}|}{\partial (\text{vec} \Sigma^{-1})'} \frac{\partial \text{vec} \Sigma^{-1}}{\partial \beta_i}, \quad (96)$$

where vec is the vec operator (Section D.1.1). The first term of Eq. (96) becomes

$$\frac{\partial \log |\Sigma^{-1}|}{\partial (\text{vec} \Sigma^{-1})'} = \frac{1}{|\Sigma^{-1}|} \frac{\partial |\Sigma^{-1}|}{\partial (\text{vec} \Sigma^{-1})'} \quad (97)$$

$$= \frac{1}{|\Sigma^{-1}|} |\Sigma^{-1}| (\text{vec} \Sigma')' \quad (98)$$

$$= (\text{vec} \Sigma')', \quad (99)$$

where we used a formula of the derivative of a determinant (Section D.2.3). The second term of Eq. (96) becomes

$$\frac{\partial \text{vec} \Sigma^{-1}}{\partial \beta_i} = \frac{\partial}{\partial \beta_i} \text{vec} \sum_{k=1}^m \mathbf{A}_k \beta_k \quad (100)$$

$$= \text{vec} \mathbf{A}_i \quad (101)$$

Substituting Eqs. (99) and (101), Eq. (96) becomes

$$\frac{\partial \log |\Sigma^{-1}|}{\partial \beta_i} = (\text{vec} \Sigma')' \text{vec} \mathbf{A}_i \quad (102)$$

$$= \text{tr} \Sigma \mathbf{A}_i \quad (103)$$

where we used a basic connection between the vec operator and the trace (Section D.2.1).

With Eqs. (95) and (103), the gradient vector given by Eq. (93) becomes

$$\frac{\partial f}{\partial \beta_i} = \text{tr} \mathbf{S} \mathbf{A}_i - \text{tr} \Sigma \mathbf{A}_i, \quad (104)$$

for $i = 1, \dots, m$.

B.2 Hessian matrix

The Hessian matrix is the derivative of the gradient vector given by Eq. (104):

$$\frac{\partial^2 f}{\partial \beta_i \partial \beta_j} = \frac{\partial}{\partial \beta_i} (\text{tr} \mathbf{A}_j \mathbf{S} - \text{tr} \Sigma \mathbf{A}_j) \quad (105)$$

$$= -\frac{\partial}{\partial \beta_i} \text{tr} \Sigma \mathbf{A}_j \quad (106)$$

$$= -\frac{\partial}{\partial \beta_i} (\text{vec} \Sigma')' \text{vec} \mathbf{A}_j \quad (107)$$

$$= -\left(\frac{\partial \text{vec} \Sigma'}{\partial \beta_i} \right)' \text{vec} \mathbf{A}_j. \quad (108)$$

We expand the first term of Eq. (108) with the elements of matrix $(\Sigma')^{-1}$ as

$$\frac{\partial \text{vec} \Sigma'}{\partial \beta_i} = \frac{\partial \text{vec} \Sigma'}{\partial (\text{vec} (\Sigma')^{-1})'} \frac{\partial \text{vec} (\Sigma')^{-1}}{\partial \beta_i}. \quad (109)$$

The first term of Eq. (109) becomes

$$\frac{\partial \text{vec} \Sigma'}{\partial (\text{vec} (\Sigma')^{-1})'} = -\left\{ [(\Sigma')^{-1}]' \right\}^{-1} \otimes [(\Sigma')^{-1}]^{-1} \quad (110)$$

$$= -\Sigma \otimes \Sigma', \quad (111)$$

where \otimes denotes the Kronecker product (Section D.1.2), and we used a formula of the derivative of an inverse matrix (Section D.2.4). The second term of Eq. (109) is represented as

$$\frac{\partial \text{vec}(\Sigma')^{-1}}{\partial \beta_i} = \frac{\partial}{\partial \beta_i} \text{vec} \left(\sum_{k=1}^m \mathbf{A}_k \beta_k \right)' \quad (112)$$

$$= \frac{\partial}{\partial \beta_i} \text{vec} \sum_{k=1}^m \mathbf{A}_k' \beta_k \quad (113)$$

$$= \text{vec} \mathbf{A}_i'. \quad (114)$$

Substituting Eqs. (111) and (114), Eq. (109) becomes

$$\frac{\partial \text{vec} \Sigma'}{\partial \beta_i} = -(\Sigma \otimes \Sigma') \text{vec} \mathbf{A}_i' \quad (115)$$

$$= -\text{vec} \Sigma' \mathbf{A}_i' \Sigma', \quad (116)$$

where we used a formula of the connection between the vec operator and the Kronecker product (Section D.2.2). Thus, the Hessian matrix given by Eq. (108) becomes

$$\frac{\partial^2 f}{\partial \beta_i \partial \beta_j} = (\text{vec} \Sigma' \mathbf{A}_i' \Sigma')' \text{vec} \mathbf{A}_j \quad (117)$$

$$= \text{tr} (\Sigma' \mathbf{A}_i' \Sigma')' \mathbf{A}_j \quad (118)$$

$$= \text{tr} \Sigma \mathbf{A}_i \Sigma \mathbf{A}_j. \quad (119)$$

C Derivation of the cost function, the gradient, and the Hessian for a fixed matrix \mathbf{A}_k given by Eq. (70)

When \mathbf{A}_k is represented as Eq. (70), we can simplify the cost function given by Eq. (13), the gradient vector given by Eq. (15), and the Hessian matrix given by Eq. (16).

C.1 Cost function

Using Eqs. (9) and (70), the first term of the cost function given by Eq. (13) becomes

$$\text{tr} \mathbf{S} \Sigma^{-1} = \text{tr} \mathbf{S} \sum_{k=1}^m \mathbf{A}_k \beta_k \quad (120)$$

$$= \text{tr} \mathbf{S} \sum_{k=1}^m (\mathbf{e}_{p(k)} \mathbf{e}_{q(k)}' + \mathbf{e}_{q(k)} \mathbf{e}_{p(k)}') \beta_k \quad (121)$$

$$= \sum_{k=1}^m \beta_k \text{tr} \mathbf{S} \mathbf{e}_{p(k)} \mathbf{e}_{q(k)}' + \sum_{k=1}^m \beta_k \text{tr} \mathbf{S} \mathbf{e}_{q(k)} \mathbf{e}_{p(k)}' \quad (122)$$

$$= \sum_{k=1}^m \beta_k (\mathbf{S})_{q(k)p(k)} + \sum_{k=1}^m \beta_k (\mathbf{S})_{p(k)q(k)} \quad (123)$$

$$= 2 \sum_{k=1}^m \beta_k (\mathbf{S})_{p(k)q(k)}, \quad (124)$$

where Eq. (144) and the symmetry of \mathbf{S} were used.

C.2 Gradient vector

Using Eq. (70), the gradient vector given by Eq. (15) becomes

$$\frac{\partial f}{\partial \beta_i} = \text{tr} \mathbf{S} \mathbf{A}_i - \text{tr} \Sigma \mathbf{A}_i \quad (125)$$

$$= \text{tr} \mathbf{S} (\mathbf{e}_{p(i)} \mathbf{e}_{q(i)}' + \mathbf{e}_{q(i)} \mathbf{e}_{p(i)}') - \text{tr} \Sigma (\mathbf{e}_{p(i)} \mathbf{e}_{q(i)}' + \mathbf{e}_{q(i)} \mathbf{e}_{p(i)}') \quad (126)$$

$$= \text{tr} \mathbf{S} \mathbf{e}_{p(i)} \mathbf{e}_{q(i)}' + \text{tr} \mathbf{S} \mathbf{e}_{q(i)} \mathbf{e}_{p(i)}' - \text{tr} \Sigma \mathbf{e}_{p(i)} \mathbf{e}_{q(i)}' - \text{tr} \Sigma \mathbf{e}_{q(i)} \mathbf{e}_{p(i)}' \quad (127)$$

$$= (\mathbf{S})_{q(i)p(i)} + (\mathbf{S})_{p(i)q(i)} - (\Sigma)_{q(i)p(i)} - (\Sigma)_{p(i)q(i)} \quad (128)$$

$$= 2(\mathbf{S})_{p(i)q(i)} - 2(\Sigma)_{p(i)q(i)}, \quad (129)$$

where Eq. (144) and the symmetry of \mathbf{S} and Σ were used.

C.3 Hessian matrix

Using Eq. (70), the Hessian matrix given by Eq. (16) becomes

$$\frac{\partial^2 f}{\partial \beta_i \partial \beta_j} = \text{tr} \Sigma \mathbf{A}_i \Sigma \mathbf{A}_j \quad (130)$$

$$= \text{tr} \left[\Sigma (\mathbf{e}_{p(i)} \mathbf{e}_{q(i)}' + \mathbf{e}_{q(i)} \mathbf{e}_{p(i)}') \Sigma (\mathbf{e}_{p(j)} \mathbf{e}_{q(j)}' + \mathbf{e}_{q(j)} \mathbf{e}_{p(j)}') \right] \quad (131)$$

$$= \text{tr} \Sigma (\mathbf{e}_{p(i)} \mathbf{e}_{q(i)}' \Sigma \mathbf{e}_{p(j)} \mathbf{e}_{q(j)}' + \mathbf{e}_{p(i)} \mathbf{e}_{q(i)}' \Sigma \mathbf{e}_{q(j)} \mathbf{e}_{p(j)}' + \mathbf{e}_{q(i)} \mathbf{e}_{p(i)}' \Sigma \mathbf{e}_{p(j)} \mathbf{e}_{q(j)}' + \mathbf{e}_{q(i)} \mathbf{e}_{p(i)}' \Sigma \mathbf{e}_{q(j)} \mathbf{e}_{p(j)}') \quad (132)$$

$$= \text{tr} \Sigma (\mathbf{e}_{p(i)} (\Sigma)_{q(i)p(j)} \mathbf{e}_{q(j)}' + \mathbf{e}_{p(i)} (\Sigma)_{q(i)q(j)} \mathbf{e}_{p(j)}' + \mathbf{e}_{q(i)} (\Sigma)_{p(i)p(j)} \mathbf{e}_{q(j)}' + \mathbf{e}_{q(i)} (\Sigma)_{p(i)q(j)} \mathbf{e}_{p(j)}') \quad (133)$$

$$= (\Sigma)_{q(i)p(j)} \text{tr} \Sigma \mathbf{e}_{p(i)} \mathbf{e}_{q(j)}' + (\Sigma)_{q(i)q(j)} \text{tr} \Sigma \mathbf{e}_{p(i)} \mathbf{e}_{p(j)}' + (\Sigma)_{p(i)p(j)} \text{tr} \Sigma \mathbf{e}_{q(i)} \mathbf{e}_{q(j)}' + (\Sigma)_{p(i)q(j)} \text{tr} \Sigma \mathbf{e}_{q(i)} \mathbf{e}_{p(j)}' \quad (134)$$

$$= (\Sigma)_{q(i)p(j)} (\Sigma)_{q(j)p(i)} + (\Sigma)_{q(i)q(j)} (\Sigma)_{p(j)p(i)} + (\Sigma)_{p(i)p(j)} (\Sigma)_{q(j)q(i)} + (\Sigma)_{p(i)q(j)} (\Sigma)_{p(j)q(i)} \quad (135)$$

$$= 2(\Sigma)_{p(i)p(j)} (\Sigma)_{q(i)q(j)} + 2(\Sigma)_{p(i)q(j)} (\Sigma)_{q(i)p(j)}, \quad (136)$$

where Eqs. (143) and (144) and the symmetry of Σ were used.

D Mathematical supplement

D.1 Definitions

D.1.1 The vec operator

Let \mathbf{A} be an $m \times n$ matrix, and let \mathbf{a}_j be its j -th column: $\mathbf{A} = (\mathbf{a}_1 \cdots \mathbf{a}_n)$; then $\text{vec} \mathbf{A}$ is defined as the $mn \times$

1 vector

$$\text{vec } \mathbf{A} = \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{pmatrix}, \quad (137)$$

D.1.2 The Kronecker product

Let \mathbf{A} be an $m \times n$ matrix, and let \mathbf{B} a $p \times q$ matrix. The Kronecker product of \mathbf{A} and \mathbf{B} is defined by the $mp \times nq$ matrix as

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{pmatrix}, \quad (138)$$

where a_{ij} denotes the (i, j) -element of \mathbf{A} .

D.2 Formulas

D.2.1 Connection between the vec operator and the trace operator

Let \mathbf{A} and \mathbf{B} be matrices of the same order. Then,

$$(\text{vec } \mathbf{A})' \text{vec } \mathbf{B} = \text{tr } \mathbf{A}'\mathbf{B} \quad (139)$$

(e.g., Magnus and Neudecker, 1999, Eq. 4 on p. 30).

D.2.2 Connection between the vec operator and the Kronecker product

Let \mathbf{A} , \mathbf{B} , and \mathbf{C} be three matrices such that the matrix product \mathbf{ABC} is defined. Then,

$$\text{vec } \mathbf{ABC} = (\mathbf{C}' \otimes \mathbf{A}) \text{vec } \mathbf{B} \quad (140)$$

(e.g., Magnus and Neudecker, 1999, Theorem 2 on p. 30).

D.2.3 Derivative of the determinant

Let \mathbf{X} be a square non-singular matrix. Then,

$$\frac{\partial |\mathbf{X}|}{\partial (\text{vec } \mathbf{X})'} = |\mathbf{X}| (\text{vec } (\mathbf{X}^{-1})')' \quad (141)$$

(e.g., Magnus and Neudecker, 1999, Table 5, p. 179).

D.2.4 Derivative of an inverse matrix

Let \mathbf{X} be a square non-singular matrix. Then,

$$\frac{\partial \mathbf{X}^{-1}}{\partial (\text{vec } \mathbf{X})'} = -(\mathbf{X}')^{-1} \otimes \mathbf{X}^{-1} \quad (142)$$

(e.g., Magnus and Neudecker, 1999, Table 7, p. 184).

D.2.5 Formulas for the unit vector

Let \mathbf{e}_i be an $n \times 1$ vector that has a value of unity in its i -th location and zeros elsewhere, and \mathbf{A} is an $n \times n$ matrix. Then,

$$\mathbf{e}_i' \mathbf{A} \mathbf{e}_j = a_{ij}, \quad (143)$$

$$\text{tr } \mathbf{A} \mathbf{e}_j \mathbf{e}_i' = a_{ij}, \quad (144)$$

where a_{ij} denotes the (i, j) -element of \mathbf{A} .

Proof Let \mathbf{a}_j be the j -th column of \mathbf{A} . Then, $\mathbf{e}_i' \mathbf{A} \mathbf{e}_j = \mathbf{e}_i' \mathbf{a}_j = a_{ij}$, which proves Eq. (143). The right-hand side of Eq. (144) is equal to that of Eq. (143), because $\text{tr } \mathbf{A} \mathbf{e}_j \mathbf{e}_i' = \text{tr } \mathbf{e}_i' \mathbf{A} \mathbf{e}_j = \mathbf{e}_i' \mathbf{A} \mathbf{e}_j$. \square

References

- Akaike H. 1974. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **AC-19**: 716–723.
- Akaike H. 1977. On entropy maximization principle. In: *Applications of Statistics, Proceedings of the symposium held at Wright State University, Dayton, Ohio, 14–18 June 1976*, Krishnaiah PR (ed). North-Holland Publishing Company: Amsterdam, pp. 27–41.
- Alves O, Robert C. 2005. Tropical Pacific Ocean model error covariances from Monte Carlo simulations. *Q. J. R. Meteorol. Soc.* **131**: 3643–3658, doi:10.1256/qj.05.113.
- Bishop CH, Hodyss D. 2007. Flow-adaptive moderation of spurious ensemble correlations and its use in ensemble-based data assimilation. *Q. J. R. Meteorol. Soc.* **133**: 2029–2044, doi:10.1002/qj.169.
- Borovikov A, Rienecker MM, Keppenne CL, Johnson GC. 2005. Multivariate error covariance estimates by Monte Carlo simulation for assimilation studies in the Pacific Ocean. *Mon. Wea. Rev.* **133**(8): 2310–2334, doi:10.1175/MWR2984.1.
- Boyd S, Vandenberghe L. 2004. *Convex optimization*. Cambridge University Press.
- Buehner M. 2005. Ensemble-derived stationary and flow-dependent background-error covariances: Evaluation in a quasi-operational nwp setting. *Q. J. R. Meteorol. Soc.* **131**: 1013–1043.
- Buehner M, Charron M. 2007. Spectral and spatial localization of background-error correlations for data assimilation. *Q. J. R. Meteorol. Soc.* **133**: 615–630.
- Chin TM, Mariano AJ, Chassignet EP. 1999. Spatial regression and multiscale approximations for sequential data assimilation in ocean models. *J. Geophys. Res.* **104**(C4): 7991–8014.
- Courtier P, Andersson E, Heckley W, Pailleux J, Vasiljević D, Hamrud M, Hollingsworth A, Rabier F, Fisher M. 1998. The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation. *Q. J. R. Meteorol. Soc.* **124**: 1783–1807.
- Courtier P, Thépaut JN, Hollingsworth A. 1994. A strategy for operational implementation of 4D-Var, using an incremental approach. *Q. J. R. Meteorol. Soc.* **120**: 1367–1387.
- Cullen MJ. 2003. Four-dimensional variational data assimilation: A new formulation of the background-error covariance matrix based on a potential-vorticity representation. *Q. J. R. Meteorol. Soc.* **129**: 2777–2796, doi:10.1256/qj.02.10.
- Daley R. 1991. *Atmospheric data analysis*, Cambridge Atmospheric and Space Science Series, vol. 2. Cambridge University Press.
- Deckmyn A, Berre L. 2005. A wavelet approach to representing background error covariances in a limited-area model. *Mon. Wea. Rev.* **133**(5): 1279–1294, doi:10.1175/MWR2929.1.
- Dempster AP. 1972. Covariance selection. *Biometrics* **28**(1): 157–175.
- Derber J, Bouttier F. 1999. A reformulation of the background error covariance in the ECMWF global data assimilation system. *Tellus* **51A**(2): 195–221.
- Derber J, Rosati A. 1989. A global oceanic data assimilation system. *J. Phys. Oceanogr.* **19**: 1333–1347.
- Desroziers G. 1997. A coordinate change for data assimilation in spherical geometry of frontal structures. *Mon. Wea. Rev.* **125**: 3030–3038.
- Diggle PJ, Ribeiro Jr PJ. 2007. *Model-based geostatistics*. Springer Series in Statistics, Springer-Verlag.

- Egbert GD, Bennett AF, Foreman MGG. 1994. TOPEX/POSEIDON tides estimated using a global inverse model. *J. Geophys. Res.* **99**(C12): 24,821–24,852.
- Evensen G. 2003. The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics* **53**: 343–367.
- Fisher M. 2003. Background error covariance modelling. In: *Seminar on Recent developments in data assimilation for atmosphere and ocean, 8–12 September 2003*. ECMWF: Shinfield Park, Reading, pp. 45–64.
- Fisher M, Andersson E. 2001. Developments in 4D-Var and Kalman filtering. Technical Memorandum 347, ECMWF, Shinfield Park, Reading.
- Frehlich R. 2006. Adaptive data assimilation including the effect of spatial variations in observation error. *Q. J. R. Meteorol. Soc.* **132**: 1225–1257, doi:10.1256/qj.05.146.
- Fu LL, Fukumori I, Miller RN. 1993. Fitting dynamic models to the Geosat sea level observations in the tropical Pacific Ocean. Part II: A linear, wind-driven model. *J. Phys. Oceanogr.* **23**: 2162–2181.
- Fukumori I. 2001. Data assimilation by models. In: *Satellite altimetry and earth sciences, International Geophysics Series*, vol. 69, Fu LL, Cazenave A (eds). Academic Press, pp. 237–265.
- Gaspari G, Cohn SE. 1999. Construction of correlation functions in two and three dimensions. *Q. J. R. Meteorol. Soc.* **125**: 723–757.
- Gaspari G, Cohn SE, Guo J, Pawson S. 2006. Construction and application of covariance functions with variable length-fields. *Q. J. R. Meteorol. Soc.* **132**: 1815–1838, doi:10.1256/qj.05.08.
- Gauthier P, Buehner M, Fillion L. 1998. Background-error statistics modelling in a 3D variational data assimilation scheme: Estimation and impact on the analyses. In: *ECMWF Workshop on diagnosis of data assimilation systems, 2–4 November 1998*. ECMWF: Shinfield Park, Reading, pp. 131–145.
- Gneiting T. 1999. Correlation functions for atmospheric data analysis. *Q. J. R. Meteorol. Soc.* **125**: 2449–2449.
- Gneiting T. 2002. Compactly supported correlation functions. *Journal of Multivariate Analysis* **83**: 493–508, doi:10.1006/jmva.2001.2056.
- Hamill TM, Whitaker JS, Snyder C. 2001. Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.* **129**: 2776–2790.
- Hollingsworth A, Lönnberg P. 1986. The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: The wind field. *Tellus* **38A**(2): 111–136.
- Houtekamer PL, Lefèvre L, Derome J, Ritchie H, Mitchell HL. 1996. A system simulation approach to ensemble prediction. *Mon. Wea. Rev.* **124**(6): 1225–1242, doi:10.1175/1520-0493(1996)124<1225:ASSATE>2.0.CO;2.
- Houtekamer PL, Mitchell HL. 2001. A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.* **129**(1): 123–137, doi:10.1175/1520-0493(2001)129<0123:ASEKFF>2.0.CO;2.
- Jackson DR, Keil M, Devenish BJ. 2008. Use of Canadian Quick covariances in the Met Office data assimilation system. *Q. J. R. Meteorol. Soc.* **134**: 1567–1582, doi:10.1002/qj.294.
- Keppenne CL, Rienecker MM, Jacob JP, Kovach R. 2008. Error covariance modeling in the GMAO ocean ensemble Kalman filter. *Mon. Wea. Rev.* **136**: 2964–2982, doi:10.1175/2007MWR2243.1.
- Konishi S, Kitagawa G. 2007. *Information criteria and statistical modeling*. Springer Series in Statistics, Springer.
- Lauritzen SL. 1996. *Graphical models, Oxford Statistical Science Series*, vol. 17. Oxford University Press.
- Lönnberg P, Hollingsworth A. 1986. The statistical structure of short-range forecast errors as determined from radiosonde data. Part II: The covariance of height and wind errors. *Tellus* **38A**(2): 137–161.
- Lorenc AC. 1992. Iterative analysis using covariance functions and filters. *Q. J. R. Meteorol. Soc.* **118**: 569–591.
- Lorenc AC. 1997. Development of an operational variational assimilation scheme. *J. Meteorol. Soc. Japan* **75**(1B): 339–346.
- Magnus JR, Neudecker H. 1999. *Matrix differential calculus with applications in statistics and econometrics, revised edition*. Wiley Series in Probability and Statistics, John Wiley & Sons.
- Menemenlis D, Chechelnitsky M. 2000. Error estimates for an ocean general circulation model from altimeter and acoustic tomography data. *Mon. Wea. Rev.* **128**(3): 763–778, doi:10.1175/1520-0493(2000)128<0763:EEFAOG>2.0.CO;2.
- Nychka D, Wikle C, Royle JA. 2002. Multiresolution models for non-stationary spatial covariance functions. *Statistical Modelling* **2**(4): 315–331, doi:10.1191/1471082x02st037oa.
- Oke PR, Allen JS, Miller RN, Egbert GD, Kosro PM. 2002. Assimilation of surface velocity data into a primitive equation coastal ocean model. *J. Geophys. Res.* **107**(C9), doi:10.1029/2000JC000511.
- Pannekoucke O, Berre L, Desroziers G. 2007. Filtering properties of wavelets for local background-error correlations. *Q. J. R. Meteorol. Soc.* **133**: 363–379, doi:10.1002/qj.33.
- Pannekoucke O, Berre L, Desroziers G. 2008. Background-error correlation length-scale estimates and their sampling statistics. *Q. J. R. Meteorol. Soc.* **134**: 497–508, doi:10.1002/qj.212.
- Parrish DF, Derber JC. 1992. The national meteorological center's spectral statistical-interpolation analysis system. *Mon. Wea. Rev.* **120**: 1747–1763, doi:10.1175/1520-0493(1992)120<1747:TNNMCSS>2.0.CO;2.
- Polavarapu S, Ren S, Rochon Y, Sankey D, Ek N, Koshyk J, Tarasick D. 2005. Data assimilation with the Canadian Middle Atmosphere Model. *Atmos.-Ocean* **43**(1): 77–100.
- Purser RJ, Wu WS, Parrish DF, Roberts NM. 2003a. Numerical aspects of the application of recursive filters to variational statistical analysis. Part I: Spatially homogeneous and isotropic Gaussian covariances. *Mon. Wea. Rev.* **131**(8): 1524–1535, doi:10.1175//1520-0493(2003)131<1524:NAOTAO>2.0.CO;2.
- Purser RJ, Wu WS, Parrish DF, Roberts NM. 2003b. Numerical aspects of the application of recursive filters to variational statistical analysis. Part II: Spatially inhomogeneous and anisotropic general covariances. *Mon. Wea. Rev.* **131**(8): 1536–1548, doi:10.1175//2543.1.
- Raynaud L, Berre L, Desroziers G. 2008. Spatial averaging of ensemble-based background-error variances. *Q. J. R. Meteorol. Soc.* **134**: 1003–1014, doi:10.1002/qj.245.
- Rhodin A, Anlauf H. 2007. Representation of inhomogeneous, non-separable covariances by sparse wavelet-transformed matrices. In: *ECMWF Workshop on Flow-dependent aspects of data assimilation, 11–13 June 2007*. ECMWF: Shinfield Park, Reading, pp. 169–183.
- Riishøjgaard LP. 1998. A direct way of specifying flow-dependent background error correlations for meteorological analysis systems. *Tellus* **50A**(1): 42–57.
- Schwarz G. 1978. Estimating the dimension of a model. *The Annals of Statistics* **6**(2): 461–464.
- Trémolet Y. 2007. Model-error estimation in 4D-Var. *Q. J. R. Meteorol. Soc.* **133**: 1267–1280, doi:10.1002/qj.94.
- Tsuchiya T, Xia Y. 2007. An extension of the standard polynomial-time primal-dual path-following algorithm to the weighted determinant maximization problem with semidefinite constraints. *Pacific Journal of Optimization* **3**(1): 165–182.
- Ueno G, Higuchi T, Kagimoto T, Hirose N. 2007. Application of the ensemble Kalman filter and smoother to a coupled atmosphere-ocean model. *SOLA* **3**(1): 5–8.
- Vandenberghe L, Boyd S, Wu SP. 1998. Determinant maximization with linear matrix inequality constraints. *SIAM Journal on Matrix Analysis and Applications* **19**(2): 499–533.
- Weaver A, Courtier P. 2001. Correlation modelling on the sphere using a generalized diffusion equation. *Q. J. R. Meteorol. Soc.* **127**: 1815–1846.

Number of neighbors	$\log \hat{\Sigma} $	$\ell(\bar{\mathbf{x}}, \hat{\beta})$	m	AIC	BIC
0	8891.2	-5886018.3	8585	11789206.7	11822424.6
4	-24.7	-4307907.8	25049	8665913.5	8762835.6
8	-758.9	-4177959.6	41234	8438387.3	8597933.8
12	-2519.3	-3866364.8	57189	7847107.6	8068388.8
20	-3266.4	-3734132.7	88726	7645717.4	7989024.6
24	-3434.6	-3704350.7	104190	7617081.5	8020223.5
28	-3856.4	-3629697.3	119682	7498758.7	7961843.9
36	-4150.0	-3577730.6	150381	7456223.2	8038092.0

Table III. Summary of covariances estimated with the covariance selection models: the number of neighbors, $\log |\hat{\Sigma}|$, the maximum likelihood ($\ell(\bar{\mathbf{x}}, \hat{\beta})$), the number of parameters ($m = \dim \beta$), and the AIC and BIC values for the models, where different numbers of neighbors are assumed.